Smart Feature Selection to enable Advanced Virtual Metrology

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Benjamin Lenz
aus Koblenz

Tübingen
2015

Tag der mündlichen Qualifikation:	12.06.2015
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Martin Bogdan
2. Berichterstatter:	Prof. Dr. Wolfgang Rosenstiel

Dedicated to everyone encouraging me to follow my own path, helping me to achieve my goals and to become the person I am.

I will contribute in shaping the future to maintain a world worth living in with the values I learned from you!

Acknowledgement

Eberhard-Karls University Tübingen At this point I would like to express my highest gratitude to all those people that enabled my research and the success of the present thesis. Special thanks are dedicated to Prof. Dr. M. Bogdan, Prof. Dr. W. Rosenstiel and Prof. Dr. W. Spruth for the opportunity to compose this dissertation including all the effort it takes for the supervision of a PhD student. I greatly appreciate their patience, all fruitful discussions and the given support along this difficult journey. Furthermore, I would like to thank M. Spüler, M. Dobler and A. Walter for their multisided contribution.

Infineon Technologies AG The present doctoral thesis was composed in collaboration with the Infineon Technologies AG. I am deeply indebted to my supervisor B. Barak for his incredible and outstanding contribution to the success of this dissertation and the support of my personal development over the past years. Also, I would like to convey my sincere gratitude to G. Schagerl for the confidence he placed in me and his remarkable support during the last years. The indefatigable support for not only the formatting of the thesis from my former diploma student and LaTeX-Jedi C. Leicht is highly appreciated. Valuable assistance, fruitful discussions and enlightening hints from U. Höckele, Dr. T. Kunstmann, K. Forster, Dr. A. Kyek, Dr. M. Brüggemann, G. Lippl, H. Lebrecht, Dr. S. Muringer, A. Zernig, Dr. U. Seidel, A. Horn, R.-M. Engel, G. Pfeifer, F. Bachl, K. Bart, A. Irlinger, L. Pfeifer, P. Grandl, S. Zimmermann, M. Frisch, K. Weinzierl, J. Zimpel, D. Polzer, J. Mührwald and B. Schrögmeier are gratefully acknowledged.

Family and Friends Along the exciting journey with all its facets from my birth to the composition of the present thesis I would like to express my highest gratitude to my love, my family and my friends. Special thanks are dedicated to my sister K. Lenz for being my tower of strength and sunrise in my most difficult moments. My parents R. Lenz and R. Lenz deserve my expression of incredible gratitude for their everlasting aid and imperturbable faith in me to follow my own path. Finally, I am very glad and proud to call S. Welsch, M. Modler, M. Broll, J. Kerler, C. Bergmann and M. Jüptner my friends. Thanks to you, my fraternity brothers, and to all the others for scientific breaks, blunt holidays and some tasty beers to clear my mind.

Executive Summary

The present dissertation enhances the research in computer science, especially state of the art Machine Learning (ML), in the field of process development in Semiconductor Manufacturing (SM) by the invention of a new Feature Selection (FS) algorithm to discover the most important equipment and context parameters for highest performance of predicting process results in a newly developed advanced Virtual Metrology (VM) system.

In complex high-mixture-low-volume SM, chips or rather silicon wafers for numerous products and technologies are manufactured on the same equipment. Process stability and control are key factors for the production of highest quality semiconductors. Advanced Process Control (APC) monitors manufacturing equipment and intervenes in the equipment control if critical states occur. Besides Run-To-Run (R2R) control and Fault Detection and Classification (FDC) new process control development activities focus on VM which predicts metrology results based on productive equipment and context data. More precisely, physical equipment parameters combined with logistical information about the manufactured product are used to predict the process result. The compulsory need for a reliable and most accurate VM system arises to imperatively reduce time and cost expensive physical metrology as well as to increase yield and stability of the manufacturing processes while concurrently minimizing economic expenditures and associated data flow. The four challenges of (1) efficiency of development and deployment of a corporate-wide VM system, (2) scalability of enterprise data storage, data traffic and computational effort, (3) knowledge discovery out of available data for future enhancements and process developments as well as (4) highest accuracy including reliability and reproducibility of the prediction results are so far not successfully mastered at the same time by any other approach.

Many ML techniques have already been investigated to build prediction models based on historical data. The outcomes are only partially satisfying in order to achieve the ambitious objectives in terms of highest accuracy resulting in tight control limits which tolerate almost no deviation from the intended process result. For optimization of prediction performance state of the art process engineering requirements lead to three criteria for assessment of the ML algorithm for the VM: outlier detection, model robustness with respect to equipment degradation over time and ever-changing manufacturing processes adapted for further development of products and technologies and finally highest prediction accuracy. It has been shown that simple regression methods fail in terms of prediction accuracy, outlier detection and model robustness while higher-sophisticated regression methods are almost able to constantly achieve these goals. Due to quite similar but still not optimal prediction performance as well as limited computational feasibility in case of numerous input parameters, the choice of superior ML regression methods does not ultimately resolve the problem. Considering the entire cycle of Knowledge Discovery in Databases including Data Mining (DM) another task appears to be crucial: FS. An optimal selection of the decisive parameters and hence reduction of the input space dimension boosts the model performance by omitting redundant as well as spurious information. Various FS algorithms exist to deal with correlated and noisy features, but each of its own is not capable to ensure that the ambitious targets for VM can be achieved in prevalent high-mixture-low-volume SM.

The objective of the present doctoral thesis is the development of a smart FS algorithm to enable a by this advanced and also newly developed VM system to comply with all imperative requirements for improved process stability and control. At first, a new Evolutionary Repetitive Backward Elimination (ERBE) FS algorithm is implemented combining the advantages of a Genetic Algorithm (GA) with Leave-One-Out (LOO) Backward Elimination as wrapper for Support Vector Regression (SVR). At second, a new high performance VM system is realized in the productive environment of High Density Plasma (HDP) Chemical Vapor Deposition (CVD) at the Infineon frontend manufacturing site Regensburg. The advanced VM system performs predictions based on three state of the art ML methods (i. e. Neural Network (NN), Decision Tree M5' (M5') & SVR) and can be deployed on many other process areas due to its generic approach and the adaptive design of the ERBE FS algorithm.

The developed ERBE algorithm for smart FS enhances the new advanced VM system by revealing evidentially the crucial features for multivariate nonlinear regression. Enabling most capable VM turns statistical sampling metrology with typically 10 % coverage of process results into a 100 % metrological process monitoring and control. Hence, misprocessed wafers can be detected instantly. Subsequent rework or earliest scrap of those wafers result in significantly increased stability of subsequent process steps and thus higher yield. An additional remarkable benefit is the reduction of production cycle time due to the possible saving of time consuming physical metrology resulting in an increase of production volume output up to 10 % in case of fab-wide implementation of the new VM system.

Zusammenfassung

Die vorliegende Dissertation erweitert die aktuelle Forschungsarbeit im Bereich der Informatik, im Besonderen den neuesten Stand der Technik hinsichtlich Maschinellem Lernen (d.h. ML) im Bereich der Prozessentwicklung in der Halbleiterindustrie, durch die Erfindung eines neuen Feature Selection (d. h. FS) Algorithmus zur Identifizierung der ausschlaggebenden Anlagen- und Kontextparameter für höchste Vorhersageleistungen der Prozessergebnisse in einem neu entwickelten fortschrittlichen System der Virtuellen Messtechnik (d. h. VM).

In der komplexen high-mixture-low-volume (Logik) Halbleiterindustrie werden Chips, beziehungsweise auf Silizium basierende Wafer, für zahlreiche Produkte und Technologien in der gleichen Anlage bearbeitet. Dabei sind Prozesskontrolle und Prozessstabilität Schlüsselfaktoren, um Halbleiter höchster Qualität zu produzieren. Fortgeschrittene Prozesskontrolle (d. h. APC) überwacht produktive Anlagen und greift im Falle kritischer Zustände in deren Steuerung ein. Neben den etablierten Systemen Run-To-Run-Control und Fault Detection and Classification fokussieren sich neue Entwicklungen der Prozesskontrolle auf VM, die die gemessenen Ergebnisse nach einem Prozessschritt an einer Anlage basierend auf den produktiven und kontextuellen Daten vorhersagt. Genauer gesagt werden physikalische Anlagenparameter mit logistischen Informationen über das gefertigte Produkt kombiniert und für die Vorhersage verwendet. Der zwingend erforderliche Bedarf für ein verlässliches und genauestes VM System entsteht, um sowohl zeitlich und finanziell sehr aufwendige physikalische Messungen zu reduzieren als auch die Ausbeute und Stabilität der Fertigungsprozesse zu erhöhen, während gleichzeitig betriebswirtschaftliche Aufwände und damit verbundene Datenflüsse minimiert werden. Bis zum heutigen Tage konnten jedoch die vier Herausforderungen (1) einer effizienten Entwicklung und Umsetzung eines unternehmensweiten VM Systems, (2) einer unternehmensweiten Skalierbarkeit der Datenspeicherung, des Datenflusses und des Rechenaufwandes, (3) einer Gewinnung von neuem Wissen aus den vorliegenden Daten für zukünftige Verbesserungen und Prozessentwicklungen sowie (4) höchster Genauigkeit einschließlich Verlässlichkeit und Reproduzierbarkeit der Vorhersageergebnisse von keinem anderen Ansatz gleichzeitig erfolgreich bewältigt werden.

Viele ML Techniken wurden bereits hinsichtlich der Eignung für die Erstellung von auf historischen Daten basierenden Vorhersagemodellen untersucht. Die Resultate sind jedoch nur teilweise befriedigend um die ehrgeizigen Ziele hinsichtlich höchster Genauigkeit zu erzielen, die in engen Kontrollgrenzen resultieren, welche so gut wie keine Abweichung vom beabsichtigten Prozessergebnis erlauben. Die Optimierung der Vorhersageleistung unter Berücksichtigung der Anforderungen modernster Prozesstechnik führt zu drei Bewertungskriterien eines ML Algorithmus für VM: Erkennung von Ausreißern, Robustheit des Modells in Bezug auf stetige Degradierung einzelner Anlagenteile mit sich ständig verändernden Fertigungsprozessen, jeweils angepasst an die Entwicklung zukünftiger Produkte und Technologien, und schließlich höchster Vorhersagegenauigkeit. Während einfache Regressionsmethoden hinsichtlich Vorhersagegenauigkeit, Ausreißererkennung und Modellrobustheit nachweislich scheitern, sind ausgeklügeltere Regressionsalgorithmen in der Lage diese Ziele großenteils zu erreichen. Aufgrund ähnlicher, je-

doch immer noch nicht optimaler Vorhersageleistung sowie begrenzter Berechenbarkeit im Falle zahlreicher Eingangsvariablen löst auch die Auswahl einer dieser fortgeschrittenen ML Algorithmen letztlich nicht das Problem. Bei Betrachtung des gesamten Zyklus der Wissensgewinnung aus Datenbanken inklusive Data-Mining erscheint eine andere Aufgabenstellung entscheidend: FS. Eine optimale Auswahl der ausschlaggebenden Parameter und die damit einhergehende Reduktion der Dimensionen der Gesamtmenge der Eingangsvariablen verbessert die Vorhersageleistung der Modelle durch den Ausschluss redundanter und störender Informationen. Zur Beherrschung von korrelierten und verrauschten Charakteristika existieren verschiedenste FS Algorithmen, wovon jeder Einzelne an sich im vorherrschenden Bereich der Halbleiterfertigung von Logikbauteilen jedoch nicht dazu geeignet ist die ambitionierten Zielsetzungen von VM zu erreichen und sicherzustellen.

Ziel und Gegenstand der vorliegenden Dissertation ist die Entwicklung eines ausgefeilten FS Algorithmus, der die Implementierung eines ebenfalls neu entwickelten fortschrittlichen VM Systems ermöglicht, das die unabdingbaren Anforderungen hinsichtlich größtmöglicher Prozesskontrolle und Prozessstabilität erfüllt. Einerseits wurde der neue ERBE FS Algorithmus implementiert, der als Wrapper für Support Vector Regression die Vorteile eines Genetischen Algorithmus (d. h. GA) mit denen der LOO Backward Elimination verbindet. Zusätzlich wurde ein hoch performantes VM System in produktiver Fertigungsumgebung im Bereich HDP CVD am Infineon Frontendstandort Regensburg realisiert. Dieses hochentwickelte VM System macht Vorhersagen basierend auf drei dem aktuellen Stand der Wissenschaft entsprechenden ML Methoden (d.h. Neuronales Netzwerk NN, Entscheidungsbaum M5' und SVR) und kann aufgrund seines generischen Ansatzes und dem adaptiven Design des ERBE FS Algorithmus in allen Prozessbereichen zur Anwendung gebracht werden.

Der neue entwickelte ERBE Algorithmus für ausgefeilte FS erweitert das neue fortschrittliche VM System erwiesenermaßen durch Identifizierung der ausschlaggebenden Charakteristika für multivariate nichtlineare Regression. Die Ermöglichung einer höchst leistungsfähigen VM erlaubt den Übergang von einer auf physikalischen Messungen basierenden statistischen Überprüfung per Stichproben mit typischerweise 10 % Abdeckung der Prozessergebnisse zu einer 100 % Überwachung aller Prozesse. Dadurch können fehlprozessierte Wafer unmittelbar erkannt werden. Darauffolgende Nacharbeit oder der frühest mögliche Verwurf dieser Wafer führt zu deutlich erhöhter Stabilität der nachfolgenden Prozessschritte und damit zu einer höheren Ausbeute. Ein weiterer beachtlicher Nutzen ergibt sich aus der Reduktion der Durchlaufzeit der Fertigung durch die Ersparnis zeitaufwendiger physikalischer Messungen, die bei fabrikweiter Umsetzung des neuen VM Systems eine Erhöhung des Produktionsvolumens um bis zu 10 % ermöglicht.

Contents

1 Introduction

A brief motivation is given to demonstrate the demand of smart Feature Selection for advanced Virtual Metrology, followed by a brief introduction of Infineon and concluded by the structuring of the present doctoral thesis.

1.1 Motivation

The innovative and capital intensive semiconductor manufacturing industry constantly strives to develop leading edge technologies and thereby to improve the complex high-end manufacturing processes every day. In order to steadily optimize the non-value-adding but indispensable physical metrology process of all intermediate products (so-called wafers) in terms of highest required accuracy for these highly complex manufacturing processes, VM actually evolved to an important research area and an expected standard operation in future. Based on historical data of logistical and process parameters, statistical models are trained to immediately predict physical metrology outcomes using actual wafer data online. Compared to significantly delayed physical measurements, on the one hand a reduction of physical metrology can be performed but on the other hand even more important is the improved quality by possible reactions and corrective actions after the process right away. Hence, the deployment for corporate-wide implementation of VM is focused. Nevertheless, the development of a corporate-wide VM implementation is nowadays still infeasible without FS. The newly developed smart FS algorithm ERBE tackles this problem in the present dissertation and provides a solution for the following challenges to enable corporate-wide and thus advanced VM. A more detailed description is provided in section 4.1.

1. Efficiency: The semiconductor manufacturing industry struggles to efficiently implement VM corporate-wide due to the lack of a generic VM approach for all equipment and processes which is essential for an economic return on invest. The focus on training prediction models incorporating only the crucial features obtained by smart FS enables a generic VM approach with high accuracy for each process.

2. Scalability: The volume and related costs to implement and deploy VM without reduction of available process parameters (i. e. 50 up to more than 10000) are corporate-wide neither scalable nor affordable in terms of data storage, data traffic, computational effort and maintainability.

3. Knowledge Discovery: Most often the pure application of various ML techniques hardly generates much information about the investigated process whereas smart FS reveals the

really crucial features yielding highest achievable accuracy which can be incorporated for further process development and enhancement in future.

4. Accuracy: High product quality requires well-controlled manufacturing which in fact can only be achieved by very accurate physical measurements and thus by comparable VM predictions often tolerating no more than 1 % deviation of the real physical metrology. The highest accuracy can be obtained by only including the most important features containing valuable information to train a statistical model whereas noisy, redundant and distracting features should be discarded. Hence, FS is required to automatically reveal the most important features and thus achieve highest possible accuracy with VM.

1.2 Infineon

Infineon Technologies is a leading Semiconductor Manufacturer in Europe headquartered in Munich, Germany. The public company was established in April 1999 as the semiconductor operations spin-off from Siemens and has a global workforce of about 30.000 employees.

"Infineon focuses on the three central challenges facing modern society: Energy Efficiency, Mobility and Security and offers semiconductors and system solutions for automotive and industrial electronics and chip card and security applications. Infineon's products stand out for their reliability, their quality excellence and their innovative and leading-edge technology in analog and mixed signal, radio frequency and power as well as embedded control. With a global presence, Infineon operates through its subsidiaries in the USA from Milpitas, California, in the Asia-Pacific region from Singapore, and in Japan from Tokyo. In the 2014 fiscal year, the company reported sales of ~4 billion Euro." [64], [65]

Within the technology sector of Infineon frontend operations new production processes are developed which are required for the manufacturing of new products designed by the different business units. The department Unit Process Development 6, responsible for APC development, supplies methods and software systems to improve monitoring and control of the manufacturing processes in order to ensure highest stability and production yield by minimizing deviations during each single process step. During the last years VM evolved as new area in APC with enormous research and development ongoing in SM. The ambitious goal to enable a capable and efficient corporate-wide advanced VM system starting with an implementation in the productive environment of HDP CVD at the Infineon frontend site Regensburg is based on the motivation outlined above.

1.3 Structure

The present dissertation is organized into ten chapters. At first (cf. chapter 1), an introduction is given with a brief motivation for the demand of VM and a short overview of the collaboration with Infineon Technologies. In chapter 2 all necessary SM and process principles are highlighted followed by fundamentals of DM, ML, FS and evaluation criteria in chapter 3. In chapter 4, the actual state of the art according to VM, SVR, Recursive Feature Elimination (RFE), GA and FS is emphasized together with a detailed description about the actual problems and challenges to implement advanced VM corporate-wide in the SM industry and the derived need for smart FS. The various requirements including the implementation for this advanced VM are outlined successively in chapter 5. Chapter 6 comprehensively describes the newly invented smart FS algorithm ERBE as the core of the present thesis. Experimental setup (cf. chapter 7) for the implementation and the ERBE algorithm follows including a comparison to other applied state-of-the-art techniques, before results (cf. chapter 8) and discussions (cf. chapter 9) are outlined. Finally, a conclusion is drawn and an outlook in chapter 10 completes the thesis.

2 Semiconductor Manufacturing

Semiconductor Manufacturing is a volatile and capital intensive industry. Production of complex nanoscale devices in hundreds of process steps requiring numerous different and expensive equipment is an enormous challenge. Highest precision, maximum yield and zero-defect quality are the major business demands to ensure efficient and profitable manufacturing of most reliable chips. High effort in research and development, innovative solutions and continuous improvement of product design and manufacturing are the main enablers in this industry for achieving the ambitious goals day by day.

Electronic circuits as composition of interconnected diodes, transistors and capacitors are designed to switch impressed current or impressed voltage in order to build up logical units for an integrated circuit. The integrated circuit itself combines an enormous network of electronic circuits based on semiconducting material (mostly silicon) and is commonly referred as a "chip". Nowadays, a tremendous range of applications exists for integrated circuits present in almost every field of today's modern society. As starting point of the value-added chain and key enabler for many products, SM is a highly technological, innovative as well as cost-driven industry and can be divided into frontend and backend production. Basically, the manufacturing of many chips on initially uniformly doped bare silicon discs, i. e. wafers, takes place in the frontend production whereas separation, bonding and packaging of these chips are performed within the backend production [174], [112], [66].

In SM frontend production, lots containing 25 identical wafers are processed in a Fabrication Plant (fab) in several main process areas: implantation, etching, deposition, diffusion, lithography, planarization and oxidation [55]. A lot is processed in each of these areas multiple times to build up a layered structure based on dielectric semiconductive and conductive films which are appropriately electrically connected. At the end, many identical chips consisting of a complex logic of electronic circuits are produced on every wafer and tested at the final wafer test for essential electrical specifications and defined functionalities [79]. After separation (sawing) the chips are electrically bonded and packaged in the SM backend for the various purposes of application. During hundreds of successive process steps the production equipment has to process the lots with highest precision. Almost no deviation can be tolerated during the production of micro- and nanoscale structures on a wafer. Therefore, process control is a key element in SM industry to ensure process stability which is crucial for product quality. During the last decades the process capability of the equipment evolved and systematically increased the opportunity to measure more and more physical process parameters (e. g. temperature, current, voltage, etc.) for all process steps. This online measurement offers a wide range of new opportunities to control every single process step. In addition to continuous equipment improvement the ex-

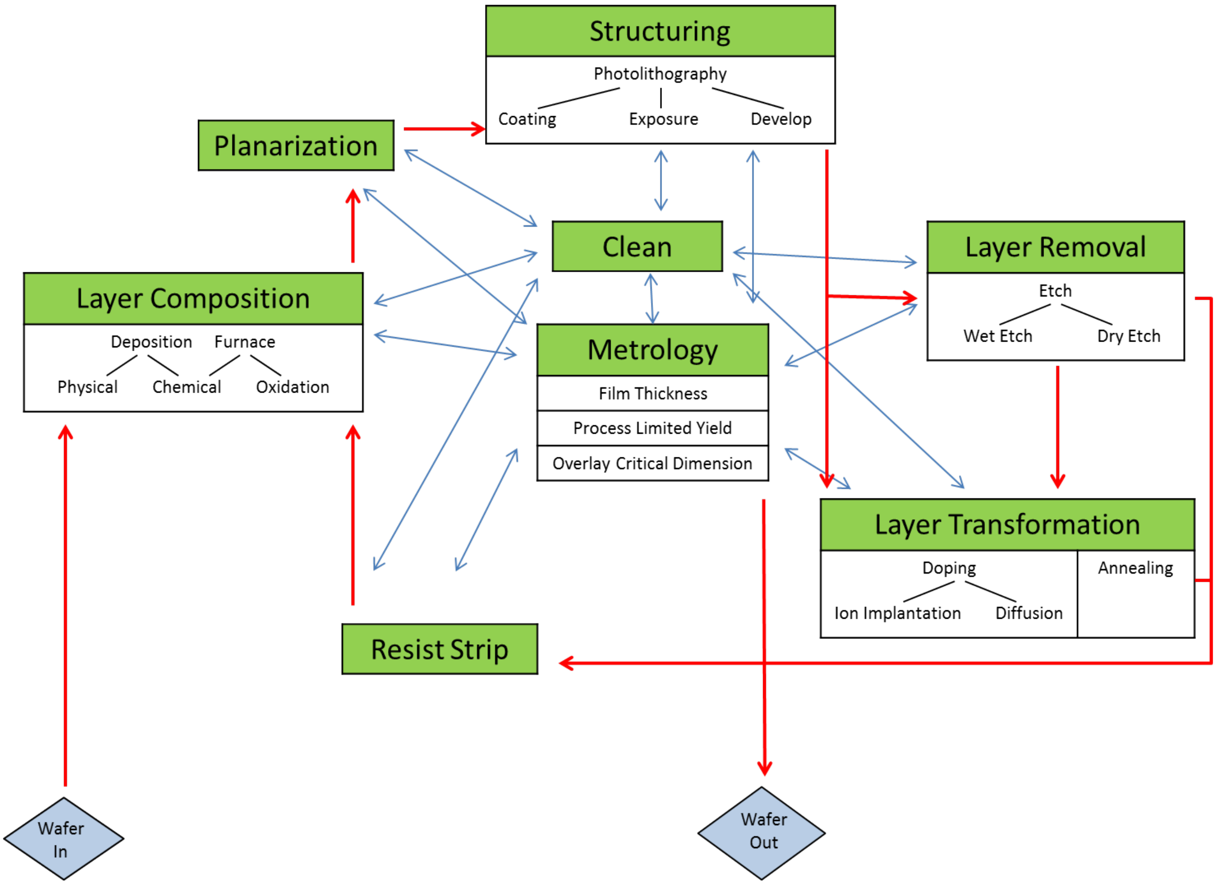


Figure 2.1: SM frontend process areas to build up a layered structure onto the wafer. Main process flow indicated by red arrows, possible intermediate process flows by blue arrows.

tended possibilities for online acquisition of high-volume process and equipment data enabled the development and implementation of APC to further improve the process capability [66], [112], [118].

Each lot is processed in all frontend process areas multiple times to build up a layered structure onto the wafers. In general, SM frontend process areas can be organized into six process sequences for altering the physical structure of a wafer. Additionally, two process sequences (i. e. Clean & Metrology) either prepare wafers for the next sequence or control the result of the last sequence. Figure 2.1 outlines the non-modifying process sequences metrology and clean (center) and the altering process sequences layer composition, planarization, structuring, layer removal, layer transformation and resist strip [174], [66]. The red and blue arrows indicate the main and intermediate process flows, respectively.

A comprehensible introduction and necessary background knowledge regarding APC and VM is outlined in the following together with a detailed description of the investigated HDP CVD and PECVD processes. More detailed information regarding the general frontend production of SM is provided in appendix A.1.

2.1 Advanced Process Control

APC evolved to a key area in semiconductor industry to monitor and to control almost all of the complex manufacturing processes. Hence, it became evident that APC is an important enabler for continuous improvement of process stability [119]. As outlined in the following, the main development areas of APC have been FDC and R2R control up to now. Recently, VM and Predictive Maintenance emerged as new promising application areas within APC.

FDC collects online data from production and metrology equipment delivered by its built-in sensors as well as by additionally integrated commercial or even self-developed sensors, yielding extended fragmentation and variety of data. Based on this collection of data, important parameters are identified and explored in order to detect any equipment or process deviation. From these parameters significant key numbers (e. g. average of O_2 pressure, standard deviation of Radio Frequency (RF) source-power) are calculated and further aggregated to obtain more condensed information. Afterwards, equipment and process specific limits are then defined to allow detection as well as classification of any undesired process and equipment failure. If a critical deviation occurs, the automated FDC system generates an appropriate online reaction which might even stop the affected production equipment. Here, the difference to standard statistical process control needs to be highlighted. The latter collects data for statistical analysis whereas APC FDC interacts with the equipment to eventually intervene the actual running process [62], [118], [120].

The second major area of APC is R2R control. Here, an adapted control algorithm is used to reduce process variability and thus to increase process capability by means of computing and applying target control settings for the specific process on the dedicated production equipment based on previously collected data. Historical data comprise parameters indicating physically measured properties of the processed wafers (e. g. thickness of a deposited dielectric layer or depth of a trench etched into the wafer surface) as well as, similar to FDC data, all related context information (e. g. equipment, process chamber, production operation, manufactured technology, product) and physical process parameters. The diversity of the context parameters further increases the fragmentation and variety of data. Many R2R controllers use an exponentially weighted moving average filter applied to input data for adjustment of the target control settings. Regular control measurements of sampled production wafers are optionally used as an additional input to R2R controllers operated in closed loop feedback mode for recalibration of the model parameters. Both, feed forward control and feed backward control are R2R methodologies to improve process performance [112], [118].

2.2 Virtual Metrology

An optimal and complete monitoring of production quality which means 100 % metrology of every single wafer after every process step is from an economic point of view by far too expensive and too time consuming making exhaustive physical metrology infeasible. Sophisticated sampling strategies to test selected wafers after specific process steps are state of the art. But for critical processes almost every wafer has to be measured to ensure the quality standards whereat the real metrology operation usually takes place with some hours delay. Thus the demand for fast and cost efficient metrology realizable by VM becomes obvious [22]. Implementation of a capable VM system enables comprehensive metrology as well as partial substitution of physical measurements by computed predictions [25], [130]. Thus, significant reduction of real metrology operations can be achieved which leads to reduced costs and product cycle time. Additionally, the production quality will be improved by instant outlier detection in case of misprocessed and defective wafers.

In case of misprocessing of a wafer at any equipment, the impact on the final product can either be detected with some hours delay, if the wafer is directly measured after the process, or not until electrical measurement in the final wafer test [90]. In the first scenario an equipment state drift between measured wafers can yield defective wafers before the next measurement. As the second scenario is more likely for typical wafer sampling rates, unnecessary waste of resources (e. g. materials and employees working time) is unavoidable. A VM system which fills the lack of physical measurement by prediction [21] [25] enables the measurement of every wafer for every process step on all capable equipment available in the fab, thus allowing significant improvement of process control as well as reduction of operational cost. Moreover, wafer-fine metrology is a requirement for real-time quality monitoring and Wafer-to-Wafer process control which is already in scope of future developments [57]. Wafer-to-Wafer process control is performed for every individual wafer in contrast to actually lot-based adjustments. Even more VM enables further analysis of wafers and of the prediction model for processes in the past. In order to implement VM, ML algorithms are trained on available historical data (i. e. process & context parameters and physical metrology results) and then applied to input data from current production to predict the associated metrology outcome [74]. In contrast to physical metrology, VM provides thereby a calculated result which is, due to the deterministic nature of the used software algorithm, highly reproducible and repeatable at any time.

The impact of VM is expected to significantly improve the effectiveness of APC in SM. So, VM was chosen by the International SEMATECH Manufacturing Initiative for the International Technology Roadmap for Semiconductors as a major research area for next generation smart factories [67]. Recent investigations estimate high benefits since corporate-wide VM implementation is expected to improve cycle time and to increase the production volume output by nearly 10 % [22]. These benefits raise the attention for VM and justify an appropriate research effort.

One use case for Knowledge Discovery and DM to improve VM is the prediction of the intermetal dielectric layer thickness in the process area of deposition. Particularly, the HDP CVD process appears to be promising for investigation of the benefits of VM (cf. subsection 5.1.1).

2.3 High Density Plasma Chemical Vapor Deposition

In HDP CVD, the reaction of a low pressure gas mixture is activated by an electrical field at low frequency (i. e. <10 MHz) inductively coupled into the process chamber by a top and a side RF coil generator for homogeneity reasons. In this RF field, gas molecules are dissociated, radicalized and energetically excited as well as positively charged by impact ionization through accelerated free electrons initially generated by collisions of the gas molecules and also by extraction from the cathode for adequate high field intensity. In the thus created localized plasma, the accelerated heavy gas ions are not fast enough to reach the cathode before the turn-over of the RF field, but enhance the activation or rather ionization of the reaction gas through collisions with other gas molecules. For a certain mixture of process gases, the impact ionization rate in the reaction gas depends on the injected RF source-power and the total gas pressure [128].

For a given RF source-power coupled into the plasma, a longer mean free path λ of the electrons at lower gas pressure facilitates the electron impact ionization and thus increases the density of charged particles in the plasma. The mean free path λ defines the distance traveled by an electron between collisions with gas molecules. In order to generate sufficient energy to ionize a gas molecule at the next impact, the mean free path has to be long enough. λ is inversely proportional to the gas pressure as a function of the density of the ionized gas molecules n and σ as the cross sectional area of these molecules (cf. equation (2.1)) [100].

$$\lambda = \frac{1}{n\sigma} \tag{2.1}$$

The localized plasma is a quasi-neutral particle system, since it contains nearly the same number of positive and negative charges. However, the negatively charged low-mass electrons, accelerated to a much higher velocity compared to the positively charged high-mass gas ions, can leave the plasma and reach the electrodes i. e. the surrounding surfaces within the process chamber, causing the build-up of a negative potential in the surface layer around the plasma. For the resulting magnitude of this so called Direct Current (DC) sheath potential, the number of positive gas ions, accelerated out of the plasma towards the process chamber dome and wafer surface (cf. subsection 2.3.1), equals in average to the number of electrons, reflected back into the plasma. This effect stabilizes the quasi-stationary and quasi-neutral state of the plasma within the process chamber [100]. A second RF coil generator induces low voltage at the wafer surface by a capacitively coupled electrical field causing a potential difference between the bulk of the plasma and the electrode, the so called DC-bias. For a given distance between the electrode and a certain reaction gas mixture and frequency of the electrical field, the DC-bias voltage U_B is a function of the process chamber pressure p and the injected DC-bias source-power W. According to [45] the DC-bias voltage is given as:

$$U_B \propto \sqrt{\frac{W}{p}}. \tag{2.2}$$

The DC-bias causes an acceleration of prior dissociated and/or ionized particles from the plasma towards the wafer surface. The energy of the positive charged gas ions before the impact on the surface can be empirically described as in [45]:

$$E_{imp} \propto \frac{U_B}{\sqrt{p}} \propto \frac{\sqrt{W}}{p}. \tag{2.3}$$

The chemical reaction resulting in the deposition of a solid film onto a wafer, situated on the electrode within the process chamber, is primarily subsisted by the radicals as well as positive gas ions generated in the plasma and made available on the wafer surface at a certain temperature. Therefore, the efficiency of the HDP CVD process, i. e. the deposition rate, increases with increasing DC-bias voltage (cf. equation (2.2)) and thus with increasing DC-bias source-power for a certain total pressure of the reaction gases in the process chamber. However, the impact energy of the positive gas ions hitting the wafer surface also increases according to (cf. equation (2.3)) with increasing DC-bias source-power. In order to avoid undesired excessive damage of the wafer surface due to the ion bombardment, the maximum DC-bias source-power injected into the process chamber needs to be limited.

Oxygen O_2 and silane SiH_4 as reaction gases build up a dielectric SiO_2 layer according to the reaction equation:

$$SiH_4 + 2\,O_2 \rightarrow SiO_2 + 2\,H_2O \tag{2.4}$$

The additionally injected partly ionized noble gas argon Ar is not comprised in the chemical reaction for the film formation but also accelerated towards the wafer surface by the DC-bias potential. The powerful impact of these positively charged Ar ions causes sputtering of the top layer of the wafer surface which is the actually growing SiO_2 film. Hence, a small fraction of the just chemically bonded SiO_2 molecules are removed again from the surface structure by the introduced energy of the Ar impact [54]. While only a small fraction of the injected argon gas is ionized and thus available for sputtering, the Deposition-Sputter ratio, especially beneficial for void-free deposition at comparable low temperatures in case of high aspect ratios for deep trench structures on the wafer surface, is the crucial setpoint to be adjusted in the HDP CVD process. The SiO_2 film deposition rate and the film removal rate due to Ar sputtering can be individually controlled according to the dependency of SiO_2 deposition on the ion/radical fluxes controlled by both the SiH_4 flow and the RF source-power and the dependency of Ar sputtering on the total ion bombardment energy (cf. equation (2.3)) controlled by the Ar flow and the DC-bias as well as RF source-power.

Furthermore, the Ar ions are responsible for the violet color of the burning plasma inside the process chamber which can be observed through a small porthole. During the ionization process in the electrical field, electrons in the atomic orbital of argon are excited to a higher state through the impact of accelerated particles. To regain a lower energetic state, the excited electrons fall back thereby emitting a photon of a characteristic energy finally causing the violet color of the HDP [100].

As a consequence of the activation of the reaction gas mixture by means of a RF fied, the resulting or rather required HDP CVD process temperature i. e. wafer temperature (for additional thermal activation of the chemical reaction) is comparably low, also allowing for processing of temperature sensitive substrates like wafers with highly doped or aluminum metalized structures. While the wafer is heated up during the HDP CVD process mainly due to the ion bombardment from the plasma induced by the DC-bias, the process relevant temperature of the wafer surface is limited by helium He gas backside cooling to an adequate value of ~ 400 °C (~ 75 °C at the backside of the wafer).

Table 2.1 illustrates typical values of the introduced HDP CVD process parameters.

Parameter Name	Unit	Value
Pressure	mTorr	<50
Ionization Rate	relative	1 %
Temperature	°C	300–450
RF source	MHz	<10
RF DC-bias	MHz	13.56
Top/Side RF power	W	~ 1000/3000
DC-bias RF power	W	~ 3000–3600

Table 2.1: Introduced HDP CVD Process Parameters [13].

The main characteristics of HDP CVD, relevant for the manufacturing process, are:

1. Void-free film deposition with high aspect ratios at comparably low temperatures of ~ 400 °C

2. Reasonable effort to ensure high purity and good quality of deposited films

3. Well defined and reproducible film composition and thickness by control of significant process parameters

2.3.1 Production Equipment

The HDP CVD production equipment dedicated to the investigation regarding the prediction of silicon dioxide layer thickness is an Applied Materials (AMAT) Centura HDP CVD mainframe platform with three out of four possible Ultima HDP CVD process chambers installed to maintain better accessibility of the central wafer handler robot chamber as shown in figure 2.2 [13]. An opened Ultima HDP CVD process chamber is outlined in more detail in figure 2.3. The multi-zone tunable inductive coupled plasma sources with the previously described top and side RF source-power coil generators are located behind the ceramic temperature controlled dome in the cover plate. The temperature of the ceramic temperature controlled dome is measured by a calibrated double thermo couple sensor. In a closed production chamber, the side RF source-power coil generator surrounds the top of the wafer surface outside of the ceramic dome and the top RF source-power coil generator stays above the initiated plasma centered above the wafer surface also outside of the ceramic dome. The multi-zone tunable gas injection is located at the center of the dome and directly below the dome as circularly arranged nozzles.

The simple symmetrically pumped chamber body contains an Ultima clean gas port to stream in the mixture of process gases whereas during a cleaning step after the productive process the clean gas (e. g. nitrogen trifluoride NF_3) is passed in via a top nozzle inlet. The cleaning process prevents excessive coating of the chamber wall due to remaining process residuals. An isolated ceramic process kit is attached to the chamber body holding the $BLUE^{TM}$ electrostatic chuck on which the wafer is processed at the chamber body center straight below the gas injection nozzles if the chamber dome is closed. The capacitively coupled DC-bias RF coil generator is connected directly to the electrostatic chuck which incorporates the He backside cooling [4].

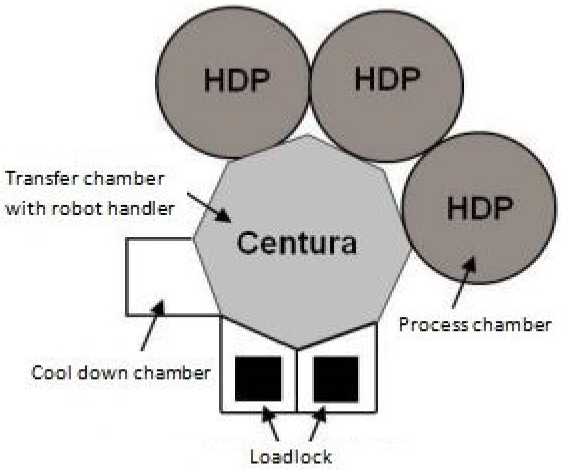


Figure 2.2: AMAT Centura HDP CVD mainframe with Ultima HDP CVD chambers [13].

2.3.2 Process Sequence

For running the HDP CVD process on productive wafers, which are stored as a lot in carriers within a wafer box for transportation in the cleanroom of the manufacturing line, the wafer carrier is taken out of the box and put into the loadlock of the AMAT Centura production equipment. After the pump down of the loadlock, the robot handler transfers the individual wafer out of the carrier through a slit valve onto jutting lift pins inside of a production ready process chamber. As soon as the robot handler is moved out of the chamber and the slit valve is closed, the lift pins retract and the wafer is laid down onto the chuck.

Subsequently the reaction process gases are supplied via calibrated mass flow controllers into the process chamber through the gas injection centered at the top and the circularly arranged nozzles at side of the wafer. The gas pressure in the process chamber, measured by a calibrated Baratron pressure gauge, is minimized and the throttle valve connected to a vacuum pump below the electrostatic chuck is fully opened. Additionally, the distance between the wafer on the chuck and the gas injection is fixed where the $BLUE^{TM}$ electrostatic chuck achieves an adjustment accuracy of $25\,\mu m$.

The chemical vapor deposition usually starts with a short deposition step as preparation prior to the main deposition step to protect sensitive metal structures on the wafer surface from degradation by sputtering due to the intensive ion bombardment in the main-deposition

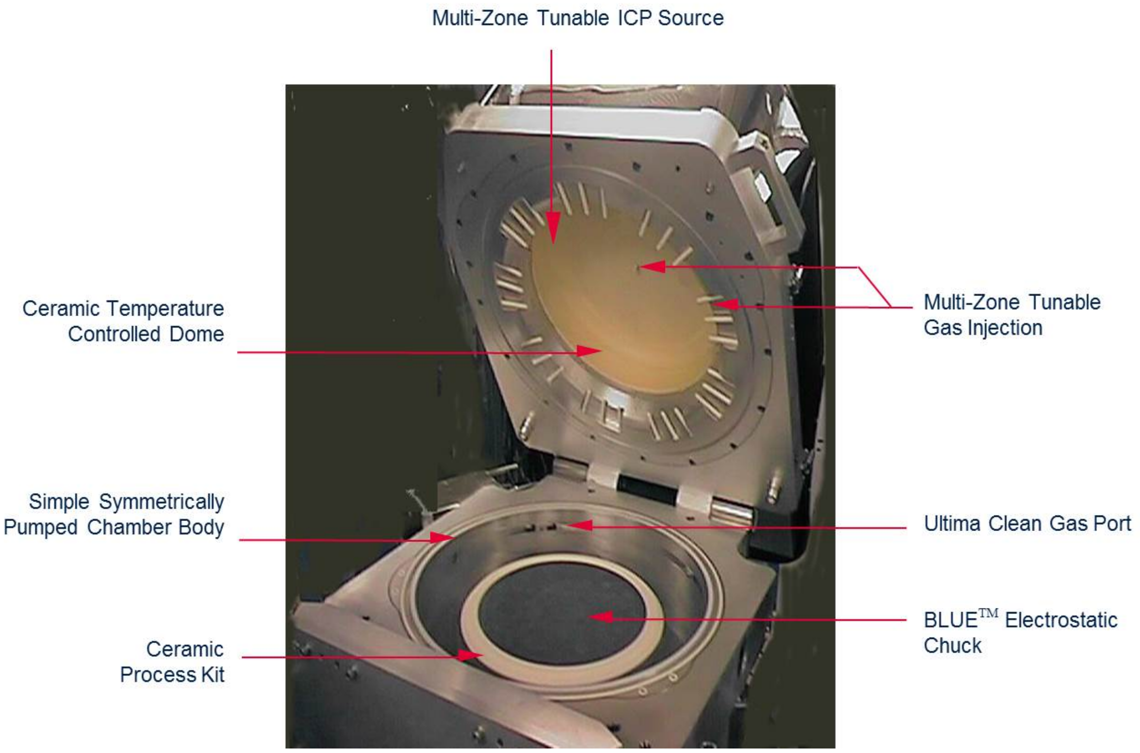

Multi-Zone Tunable ICP Source
Ceramic Temperature Controlled Dome
Multi-Zone Tunable Gas Injection
Simple Symmetrically Pumped Chamber Body
Ultima Clean Gas Port
BLUE™ Electrostatic Chuck
Ceramic Process Kit

Figure 2.3: An opened AMAT Ultima HDP CVD chamber [13].

step. Hence, a so called SiO_2 liner sub-layer is deposited without application of the DC-bias source-power [13].

Before the subsequent main deposition step of the silicon dioxide layer can be started, the plasma strike has to be performed at a lower chamber pressure of ~30 mTorr in the absence of silane and oxygen to start heating the wafer to process temperature. Thus, only Ar gas is present and the preset top RF source-power is switched on to initiate the ionization avalanche and thus to strike the plasma between the gas injection and the surface of the wafer. Additionally, the first process gas O_2 is injected into the chamber and adjusted. Now, the He gas is streamed into the vacuity between wafer and electrostatic chuck while a He flow check is performed to ensure a reliable He backside cooling of the wafer necessary for a controlled heat up to the desired process temperature. In the next step SiH_4 gas is injected into the process chamber and all process gas flows (i. e. Ar, O_2 & SiH_4) are adjusted and stabilized at the process recipe setpoints together with the wafer temperature via the He cooling. Afterwards, the DC-bias RF source-power is switched on and ramped up yielding the required DC-bias voltage which controls the deposition rate of the SiO_2 layer formation on the wafer surface. As soon as the duration of the main deposition step has reached the process recipe setpoint for achieving the required SiO_2 layer thickness, the silane flow is stopped and the DC-bias source-power is ramped down. Finally, all remaining process gases (i. e. O_2 & Ar) as well as the He cooling are switched off and the process chamber is pumped down via the fully opened throttle valve. Subsequently, the wafer is transferred by the robot handler into the cool down chamber and the process chamber is ready for the cleaning step.

After cooling down close to ambient temperature, the robot handler finally transfers the processed wafer back into the carrier within the loadlock. In the meantime, the cleaning step (using a highly reactive gas NF_3) is running to remove process residues from the chamber walls to avoid excessive buildup.

All equipment are regularly subject to major maintenance activities typically occurring within an annual or semi-annual time period. During this maintenance the equipment is set offline and all production chambers are opened, partially disassembled and thoroughly cleaned as well as controlled or repaired if required. Following the reinstallation and startup of the equipment various tests are performed on non-productive wafers to readjust the most important equipment settings including some process parameters (i. e. features) and finally resume manufacturing of the high product mix within the very small defined process windows including tight control limits in SM. No preferable or overall optimal specific value exists for these adjusted process parameters resulting in a natural variation of these process parameters due to that fact that for economic reasons not all but only broken or heavily degraded spare parts are necessarily changed during the maintenance resulting in different states of degradation of various other spare parts of the equipment yielding different value settings of process parameters for these spare parts depending on their status. Hence, these few adjusted process parameters (even neglecting the way bigger part of all other just measured but highly interrelated process parameters) show a high variety of value ranges and distributions but still yielding a target outcome within the same defined process windows. Even on top of this complexity ever changing recipes including their individual settings for high mixture SM exacerbate any intended derivation and understanding of principles of cause and effect of changed process parameters and resulted target variation for all highly complex SM processes (i. e. especially the superpositioned HDP CVD process). Whereas for highly complex physical processes it is even for very educated and experienced process experts only feasible to concurrently adjust some few parameters at the same time and with it implicating the effects on the final target, a nonlinear multivariate FS method can handle and analyze a substantial quantity of interrelated input features.

2.3.3 Physical Metrology: Optical Layer Thickness Measurement

As a very essential process, physical metrology of the obtained process results (e. g. thickness of deposited layers) is performed after each process area to detect irregularities and failures as early as possible within the process chain. These measurements prevent unnecessary degradation of production equipment, waste of valuable materials and labor time as well as increase product quality but at the expense of product cycle time. Due to the sake of an enormous amount of processed wafers, comprehensive physical metrology (cf. section 2.2) and thus complete production monitoring is simply infeasible [90]. Therefore, measurements are only operated for statistically sampled wafers depending on rules based on the associated logistical specification of the corresponding lot. In order to ensure highest possible production quality by means of statistical process control, the measurement results are continuously monitored by an appropriate software tool including statistically calculated control limits and process and/or product dependent specification limits. Control limit violations trigger remeasurement of the affected wafers or even the whole lot as well as inspection of the production equipment as appropriate. In case of confirmed specification limit violations, the faulty wafers are scrapped immediately and thus excluded from any further processing [66].

The metrological verification of the HDP CVD process result, in terms of deposition of the silicon dioxide film with a thickness within the allowed tolerance limits close to the target value of the specific inter-metal dielectric process, is done by an optical film thickness measurement for each sampled wafer on an Opti-Probe metrology equipment. After each HDP CVD process, there are several wafers measured for every production lot of the considered product technology type, at least one wafer for each of the three process chambers of the dedicated production equipment. The SiO_2 layer thickness is measured for each sampled wafer at nine measurement points evenly distributed over the wafer to avoid local dominations whereupon the mean value calculated from these measurements result in the average layer thickness of the silicon dioxide layer deposited onto the wafer. The metrology via optical layer thickness measurement performed by the Opti-Probe 3290 is very accurate and precise with only minor variations (0.4 % accuracy, 0.05 % precision & 0.01 % reliability – cf. subsection 5.2.3) [161].

Due to the deposition of silicon dioxide for building up both the underlying pre-deposition liner and the main-deposition layer resulting in a very homogeneous interface between the two layers, the optical measurement can only deliver the total thickness of the two-layer stack. For individual process control the liner thickness is regularly measured after deposition onto a specific test wafer according to the pre-deposition step of the associated productive recipe.

2.4 Plasma Enhanced Chemical Vapor Deposition

In Plasma Enhanced Chemical Vapor Deposition (PECVD), the reaction gas mixture is activated by an electrical field at RF of 13.56 MHz capacitively coupled into the process chamber. Similar to the HDP CVD process, in the RF-field gas molecules are dissociated, radicalized and energetically excited as well as positively charged by impact ionization through accelerated free electrons initially generated by collisions of the gas molecules and for sufficiently high field intensity additionally by extraction from the cathode. The accelerated heavy gas ions are not fast enough to reach the cathode before the turn-over of the RF field, but enhance the ionization of the reaction gas through collisions with other gas molecules. For a certain mixture of process gases, the impact ionization rate in the reaction gas depends on the injected RF-power and the total gas pressure [45]. As a consequence of the activation of the reaction gas mixture by means of a RF-plasma, the required process temperature (i. e. wafer temperature) is comparably low in a range of typically 200 °C to 500 °C. An important aspect of this technique is the well-defined and reproducible composition and thickness of the deposited film achievable with reasonable effort by control of the significant process parameters [55]. The PECVD metal passivation process comprises the primary deposition of a Silicon Oxide SiO_2 base layer onto a metal layer stack and the subsequent deposition of a Silicon Nitride Si_3N_4 cap layer as shown in figure 2.4:

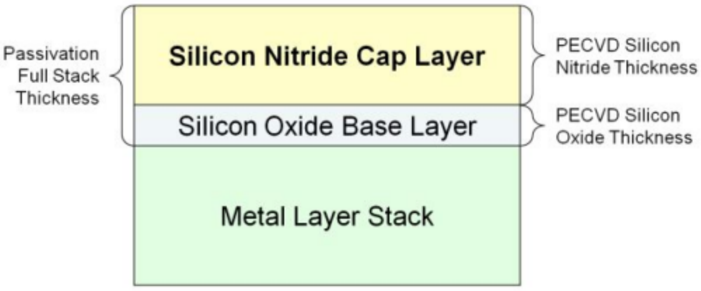

Passivation Full Stack Thickness
Silicon Nitride Cap Layer
Silicon Oxide Base Layer
Metal Layer Stack
PECVD Silicon Nitride Thickness
PECVD Silicon Oxide Thickness

Figure 2.4: Metal Passivation Layer Structure: Silicon Nitride cap layer and Silicon Oxide base layer deposited in a PECVD process sequence for passivation of the underlying metal layer stack.

Starting the PECVD process, two wafers are transferred into each of the three twin-chambers of the production equipment AMAT Producer which is able to concurrently process two wafers within a single process chamber situated twin-chamber next to each other on identical electrically heated ceramic chucks. The temperature of both wafers as well as the flow of the process gases are adjusted to the required recipe set points whereat the latter are supplied by a showerhead above the wafers. The pressure in the process chamber is controlled by a throttle valve connected to a vacuum pump. The deposition step of the first PECVD process for depositing the Silicon Oxide base layer is started by injecting the RF-power into the twin-chamber generating the required plasma between the showerhead and the surface of the wafer. As soon as the required process time of this deposition step elapsed, RF-power and all process gases are switched off. Subsequently, the process chamber is purged and the second PECVD process is started. After the deposition of the Silicon Nitride cap layer, the process chamber is pumped down again and both wafers are transferred to a cool down chamber from where they are finally moved back to the carrier in the loadlock of the equipment.

After the PECVD metal passivation process sequence, several wafers are selected and measured for every production lot depending on the individual product technology type, at least one wafer for each of the three process chambers. For each sampled wafer, the thickness of both the silicon nitride and the silicon oxide layer in the dual-layer metal passivation stack is individually measured specifically for each basic design type at several measurement points evenly distributed over the wafer. As an indicator for the quality of each measurement result the goodness of fit is used. The mean values calculated from these individual measurements are the average thickness of the Silicon Oxide base and the Silicon Nitride cap layer deposited onto the wafer. Based on this measured layer thickness, the deposition time for the next lot of wafers of the same design type is calculated by a R2R controller running for each process chamber in closed loop mode on the PECVD production equipment.

Compared to the previously described HDP CVD process both deposition processes significantly differ in regard to the following aspects:

1. *Chambers*: All chambers installed on the AMAT Centura process only a single wafer at a

time compared to the twin-chambers of the AMAT Producer where possible interactions or even exchange of material between both concurrently processed wafers cannot be excluded.

2. *Deposition Sequence*: At first, the SiO_2 base layer is deposited onto a metal layer stack followed by the Si_3N_4 cap layer during the PECVD process on AMAT Producer compared to the deposition of a single Silicon Oxide layer onto a thin SiO_2 liner during the HDP CVD process on AMAT Centura.

3. *Deposited Materials*: SiO_2 and Si_3N_4 are successively deposited in the PECVD process on AMAT Producer compared to pure SiO_2 in the HDP CVD process on AMAT Centura.

4. *Superposition of deposition and sputtering*: The PECVD process on the AMAT Producer is a conventional deposition process of SiO_2 and subsequently Si_3N_4. In contrast, the HDP CVD process consists of the superposition of the deposition of SiO_2 and the concurrent sputtering of the deposited SiO_2 by Ar at the wafer surface.

Summary: The present chapter introduces the area of VM and accurately describes the manufacturing processes HDP CVD and PECVD as demonstrative use cases for FS and VM in SM. This process knowledge is necessary to gain insight into the complexity of the environment of VM in SM determining the conducted experiments and the achieved results as well as to understand the significance of the newly invented smart FS algorithm and the developed advanced VM system based on this new FS method. The following chapter provides essential fundamentals for these new approaches.

3 Fundamentals

Principles of Data Mining, Machine Learning, Feature Selection, Support Vector Regression and Genetic Algorithm will be reviewed together with applicable evaluation criteria for Virtual Metrology.

3.1 Data Mining and Knowledge Discovery

The common approach of Knowledge Discovery in Databases was adapted and enhanced with differently defined phases and tasks for industrial application as the CRoss Industrial Standard Process for Data Mining (CRISP-DM) [20], [28]. The fundamental CRISP-DM phases are outlined in the following and the VM specific adaption at Infineon is given in subsection 4.2.2.

At first, business understanding is conducted. Afterwards, the iterative core phase is carried out. The core encompasses the three phases data understanding, Data Preparation (DP) and Modeling. The modeling phase substitutes the DM step of the Knowledge Discovery in Databases approach. If a suitable model is found, it will be tested and evaluated. Now, a terminal deployment phase is aimed for but in case of unsatisfactory results either the core phase is conducted again or in the worst case the business understanding phase needs to be reconsidered [20]. Figure 3.1 illustrates the CRISP-DM model.

Business Understanding

The initial phase of business understanding sharpens the comprehension of the final goal to be achieved. It is an important step, because the awareness of the exact definition and specification of the aimed goals will speed up the entire process by focusing on the important actions all the time. Sometimes CRISP-DM already ends here if the possibilities of Knowledge Discovery do not meet the expectations of the stakeholders. Negligence of business understanding can lead to conflicts between stakeholders and Data Miners at the end of the process. Precise monetary goals expected by the stakeholders are often hard to estimate in advance of the overall process of Knowledge Discovery and therefore point out the importance of business understanding.

Data Understanding

Data understanding as second phase is the first iterative core phase. Usually the core phases are executed several times according to Knowledge Discovery in Databases due to new available knowledge arising from other phases. Here, a detailed overview of available data and the corresponding characteristics is created. Before selecting any modeling algorithm, data shall be analyzed unprejudiced. Particular tasks are:

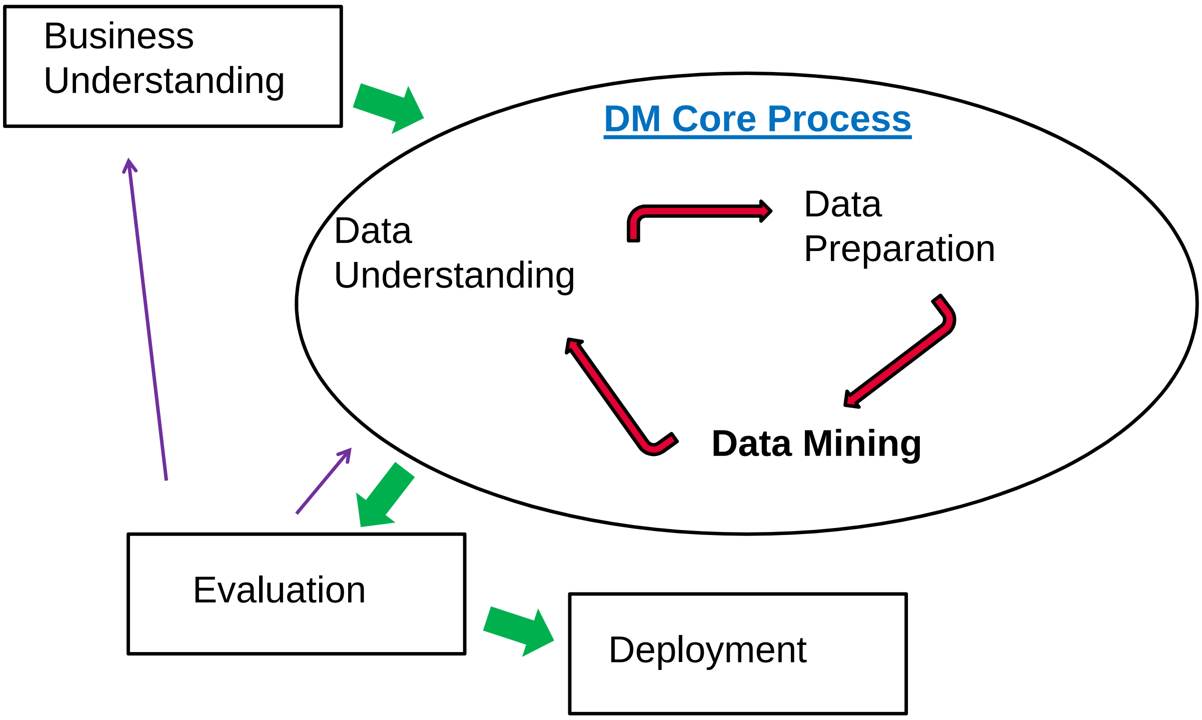


Figure 3.1: The CRISP-DM model [20]. The wide green arrows emphasize the general DM sequence while the thin purple arrows indicate necessary checks within the evaluation phase. Inside the DM core process the red arrows show the cycle of the iterative DM phases.

- *Data availability* explores all possible data sources in terms of any valuable information.
- *Data quality* examines the available data for further usage (e. g. missing data, wrong data).
- *Correlation analysis* inspects the correlation between any attributes/features for important characteristics (e. g. mean, variance).
- *Cluster analysis* studies available data regarding any present clusters to create feature or instance subsets.
- *Exploration* verifies or rejects any initially stated hypotheses or expectations.

Data Preparation

The third phase as second part of the core procedure investigates the possibilities to make data accessible for future modeling. Every algorithm can only produce accurate, robust and reliable models and predictions if the basis of data, on which it is trained, is of high quality. Noisy data, missing values and mixed subsets can degrade the accuracy, precision, specificity and sensitivity of every modeling technique to a level at which it is not acceptable anymore. Therefore, the subsequent tasks are essential aspects according to Knowledge Discovery in Databases which are often underestimated:

- *Data Formatting* can convert data between logical (binary), nominal (categorical), ordinal (ranked), interval (numerically ranked) and ratio (numerically ranked on defined scale) data.
- *Data Set Compilation* specifies whether a training dataset is recorded from a fixed dataset, from a dynamic dataset (e. g. increasing by appending new instances) or an adaptive dataset for a fixed time range (Moving Window (MW)).
- *Feature Translation* can build up new features by combining or converting existing features. Also, normalization and standardization are common actions. Moreover, conversion to different scales, convolution of various features and aggregation of features to new ones are further actions which can produce improved results.
- *Feature Transformation* aims to diminish the number of variables by dimensionality reduction methods transforming data into other representations to simplify processing or calculation.
- *Instance Selection* defines the basic inclusion or exclusion of available data instances. In addition, missing values can be excluded or replaced (e. g. by means), outliers can also be included or excluded and the global dataset can be reduced or weighted according to these considerations.
- *Feature Selection* separates features containing valuable information from those contributing mainly noise or no information (e. g. static features). The feature set can be reduced by expert advice or dedicated algorithms. In the scope of the present thesis a new FS algorithm (cf. section 6.5) is developed to reduce the initial amount of features to overcome the stated challenges and problems (cf. section 1.1, section 4.1).

Modeling

In the last core phase, different modeling techniques are investigated in terms of applicability. A huge and steadily growing variety of algorithms is available accompanied by frequent improvements. These algorithms can be classified regarding various characteristics (e. g. heuristic vs. deterministic). Important tasks are:

- *Evaluation of DM algorithms* summarizes the requirements (e. g. nominal input data), compares the advantages and drawbacks and highlights accuracy, robustness and precision.
- *Modeling of the selected DM algorithms* explores specific boundary conditions and optimizes individual model parameters to avoid overfitting while assuring precise predictions. DM algorithms can detect and recognize characteristics, structures and patterns included in available data which are hidden for human recognition and overlaid by physical or statistical noise.

- *Combination of the selected DM algorithms* supplies eventually additional benefit by exploitation of the individual advantages.

Evaluation

In the fifth phase, the core process and the initially sharpened Business Understanding are evaluated. It differs from the testing of specific DM algorithms as it is not restricted to certain goals (e. g. accuracy) but evaluates the applicability of the complete system. Sometimes, goals defined in the first phase or the DM process need to be adjusted or the entire approach can turn out to be infeasible for the intended goals. Important tasks of this phase are:

- Final tests of the elaborated DM system are performed with so far unused representative datasets. Depending on the DM technique different test methods are appropriate (e. g. ROC curves, Recall-Precision, numeric measurements (cf. section 3.6)).
- The initially stated goals are checked and the DM core process is reviewed and executed again in case of expected improvements/enhancements.
- Definition of next steps to deploy the extracted knowledge.

Deployment

The last phase focuses on the deployment of the developed CRISP-DM system either as software integrated into another system or as stand-alone application for future use. The following tasks are essential:

- Creation of a deployment plan.
- Compilation of a final report and accomplishment of a project review.

3.2 Machine Learning

ML needs to deal with data changing over a period of time while still preserving prediction performance. Hence, incrementally growing datasets and/or MW datasets can be used for model training. Another well-known challenge for ML is the overfitting-avoidance bias and bias-variance trade-off between overfitting (i. e. the target function perfectly fits to the characteristics in the training dataset) and generalization (i. e. deviations of the target function are tolerated to ensure prediction with sufficient accuracy also for unknown but similar future instances). Also, the demand for computational efficient algorithms has to be addressed in the field of ML [175], [52].

On the one hand, ML techniques can be categorized into Supervised, Semi-Supervised and Unsupervised Learning. On the other hand, ML methods are often classified as Cluster Analysis, Classification and Regression algorithms. The latter is used here and detailed in the following.

3.2.1 Methodology

Supervised Learning defines the ML task to find a useful function or pattern in labeled data. Hence, the ML method can be trained due to a rewarding of the method whether the label was predicted correctly or not. On the other hand, unsupervised learning attempts to find a not obvious function or pattern in unlabeled data whereas the solution cannot be evaluated due to the lack of labels [52].

Cluster Analysis

Cluster Analysis is an unsupervised learning approach to group objects/instances into subsets or clusters. Similar objects shall be more closely related to each other. Instances are characterized by its variables/features or relation to others. Furthermore, clusters can be ranked hierarchically. The choice of distance or dissimilarity measure is an important and essential task during cluster analysis. Analogies are present to the definition of a loss function for prediction in supervised learning [52].

Classification

Classification is a learning approach to predict whether the actual instance belongs to one of two or more classes based on characteristics of all features. The prediction result can be manifold in case of multi-class prediction and the evaluation of the result is binary and either true if the predicted class was correct or false if it was incorrect. The success rate of a technique can be evaluated in various ways (e. g. confusion matrix). Input data depends on the algorithm but can be of any format (logical, nominal, ordinal, interval, ratio) or transformed into an appropriate format. Some methods extend the classification to a regression approach by discretizing the target classes into an almost continuous multi target range [175].

Regression

Regression is also a supervised learning approach whereas the outcome is a numeric value rather than a category. Just as classification the prediction is based on characteristics of all features. Similarly, input data depend on the algorithm and can be of any format or being transformed as for classification. The predicted target can be multidimensional [175] and the outcome usually defines a continuous range of values. For instance, the prediction target of the present thesis (the deposited Layer Thickness) takes such continuous as well as positive values.

3.2.2 Regression Techniques

A brief overview of several ML regression techniques is provided below. M5’ decision trees and Back Propagation Neural Networks (BPNN) are investigated in terms of applicability for VM and used at Infineon for comparison of the prediction accuracy of SVR [87] which is described in more detail in section 3.4.

Decision Tree

The decision tree learning technique can be used for classification and regression. Initially a tree is grown where at each node from the root to the leaves a decision or split is made. At each node all remaining data are divided into two or more sub-partitions according to the most significant differentiator of any feature. In general, the feature stays in the global feature set and also several splits could be performed with respect to the same feature until a leaf is reached. The growing of the tree stops if any defined stop criterion is met (e.g. sufficient prediction accuracy by remaining instances in a leaf or minimal number of remaining instances (data) in a leaf). As second major part the built tree is pruned and optimized according to defined criteria to minimize the final decision tree. CART, ID3 and C4.5 are popular examples. M5’ is another CART-based model tree. The basic approach is to minimize the standard deviation at each split. In each leaf a linear regression is performed to obtain a smooth regression [171].

Neural Network

Artificial Neural Networks (i. e. NNs) are composed of an interconnected net of nodes (artificial neurons) inspired by biological NN. Basically, a layer of input neurons represent the input variables and a layer of output neurons represent the target variables. In between hidden layers can be established. Each node (neuron) of a hidden or output layer can use an arbitrary amount of input variables from the previous layer (input or hidden layer). The global function approximated by the NN is a composition of other functions inside the neurons of the hidden layers. Usually, neurons of the hidden layer consist of an activation function and a transfer function. The former decides if the neuron is active (used) and its result is used for further computation. The latter computes the output of the neuron. Feed forward NN describes the approach to prohibit feedback to neurons of the same or previous layer. A BPNN is trained by starting at the target proceeding layer by layer backwards to the first hidden layer. Many NN are based on the Levenberg-Marquardt algorithm [11], [52].

Support Vector Regression

SVR evolved as an extension of Support Vector Machine (SVM). SVM is an instance-based ML method. It aims to find a small number of so-called support vectors which are important boundary instances in the dataset. The approximated discriminant function aims to separate the support vectors and to maximize the margin in between [175]. A detailed description of SVR is given in section 3.4.

3.3 Feature Selection

FS is a crucial step in DP within the DM core process. Elimination of noisy data and redundant or irrelevant features (i. e. variables) yields several important advantages:

1. Improvement of prediction performance and accuracy [48], [81].

2. Speed-up and cost efficiency of a prediction system due to less computational overhead [48], [150], [72].

3. Fundamental knowledge discovery if only features remain that contain the essential information about the current application [48], [72].

4. Facilitates or enables data visualization since hundreds to thousands and even more features are not manageable. Nevertheless, conclusions have to be drawn carefully since missing features might lead to wrong inference.

The process of FS is often described as heuristic state space search where each state is modeled by a feature subset [83]. The entire state space is huge and the problem to find the best state is combinatorial an intractable, NP-hard problem which also applies to the problem of finding the feature subset that generates the best result with highest prediction accuracy for a specified induction technique [150], [115], [49]. Furthermore, according to Occam's razor, a smaller feature subset is favored in a future VM prediction system to achieve simplicity as well as scalability and minimize necessary data traffic, database storage and computational effort.

The challenge of feature selection comprises one of the two goals which are usually not coincident and compete with each other as direct objective optimization [48], [43]:

1. Find a feature subset that minimizes the prediction error.

2. Minimize the number of features for a tolerated prediction error.

Basically, the overall process of FS starts with the generation of a feature subset. In terms of Stepwise Selection (SS) or backward elimination a single feature is added to an empty pool or removed from the entire feature pool, respectively. Otherwise, an initial feature subset is generated by a defined event. Subsequently, the performance of the feature subset for a chosen ML technique is evaluated by a defined evaluation method. If one or several stopping criteria are met the feature subset will be fixed for subsequent training and prediction. A final validation of the result concludes the approach [150].

FS methods can be divided into three classes:

1. **Filters**: The process of FS is performed at first and independent of the succeeding ML method. As a preprocessing step, features are selected by a defined scoring function which is usually independent of the performance measurement of the later learning algorithm.

2. **Wrappers**: A wrapper approach incorporates the ML method into the DM process. The learning algorithm is trained on various feature subsets. The optimization and FS is assessed with the final evaluation criteria.

3. **Embedded Methods**: In contrast to wrappers an embedded method involves the FS in the training process [48].

3.3.1 Filters

A filter approach ranks features by a scoring function according to general characteristics prior to any prediction. It is independent from the subsequent prediction technique [150]. Filters work well for individual and independent variables. Thus, considering independence or orthogonality assumptions, a filter may be an optimal choice. Advantages of filters are simplicity, scalability to very high dimensional datasets and good empirical success [139]. Fairly simple scoring functions are computational efficient and independent of the prediction algorithm (i. e. classification or regression). Filters are quite robust against overfitting due to their statistical and non-learning nature with less introduced bias [52], [48]. A mayor disadvantage is the complete disjunction of the filter from the performance of the following prediction algorithm [177]. The selected feature subset is not linked to the prediction technique which can indeed learn underlying characteristics in addition to independent statistical analysis [81]. Another drawback is the linear filter approach in case of a nonlinear environment. Also, most filters are univariate resulting in worse performance compared to other FS techniques if feature dependencies are present in the dataset [139]. A common approach is to apply a filter prior to a wrapper to overcome complexity and reduce the initial feature set for the wrapper [48]. According to [48] examples of filter approaches are:

- Fisher's Linear Discriminant Analysis (LDA)
- The Pearson Correlation Coefficient
- The RELIEF Algorithm (cf. section 4.3)
- Individual Feature Scoring
- Mutual Information between each feature and the target

3.3.2 Wrappers

A wrapper approach comprises the ML method as black box and 'wraps' the deterministic or randomized search method around the prediction algorithm. The prediction performance of the learning machine is assessed for various feature subsets as input. Hence, it searches for the best suited feature subset for this specific ML method. It is not necessarily the overall best feature subset and other learning techniques might be optimal for other subsets [48], [81], [115]. Incorporation of the ML method as black box makes wrappers more universal in their application. Their advantages are first of all the interaction between the prediction algorithm and the search for the best feature subset and as second the incorporation of hidden interdependencies between various features [139], [177] and as third the often superior prediction performance [72]. As possible drawbacks to consider, wrappers sometimes tend to be computational expensive because they need to retrain the model for each feature subset and they have a higher susceptibility to overfit the model to the training dataset [48], [72], [81], [115], [139]. In general and according to [139], several search strategies exist:

1. **Greedy Search Strategies**
   a) **Forward/Stepwise Selection**: Features are added stepwise or in a bunch to the initially empty feature subset.

   b) **Backward Elimination**: Features are removed stepwise or in a bunch from the initially complete feature set.

   c) **Hill-Climbing**: A combination of forward selection and backward elimination to find an optimum.

2. **Advanced Search Strategies**
   a) **Best First**: The most promising feature (e. g. as result from a filter/feature ranking approach) is added to the final feature subset. Until a defined threshold is met more promising features are added. The final feature subset is evaluated and inherited if it yields a better prediction performance as the initial feature subset [81].

   b) **Branch-and-Bound**: Like a tree, the search space is explored in the direction of all features with stepwise increasing the number of features (branch). If the explored branch does not meet a certain threshold after some branches are made, the failing branch is cut off and will not be considered anymore (bound) [121].

   c) **Simulated Annealing**: Inspired from material science, the method optimizes the feature subset by slowly decreasing (cooling) the probability to accept other solutions which degrade the prediction performance [35]. This and the following three techniques are metaheuristic algorithms not guaranteeing the optimal solution but providing a universal search with few or no assumptions.

   d) **Genetic Algorithm**: The features are decoded as binary vector. Applying mutation and crossover operations to the actual binary feature representation generates different feature subsets. Starting with an initial population, the technique aims to optimize the binary represented feature subset in each generation by inheriting the fittest features for populating the next generation by the stated operations [117], [144] (cf. section 3.5 & section 6.4). Compared to simulated annealing GAs evaluate more than one candidate in each generation and are stated to perform equal or slightly better if more time is available (slow starter) and slightly inferior vice versa (simulated annealing as quick starter) [109].

   e) **Particle Swarm Optimization**: Inspired by flock of birds, the method optimizes a moving population of candidate solutions (i. e. feature subsets) around the search space along position and velocity vectors. The candidates do not compete as in GA but cooperate to find the best solution. Each contributing candidate is constantly influenced by its individual and the global swarm best position and speed. Similar to simulated annealing the method initially focuses on exploration for the area around the best solution and later exploitation and optimization of this solution [17], [7], [73].

f) **Ant Colony Optimization**: Inspired by ant colonies, the technique randomly optimizes the path through a graph reinforcing already successfully traveled trails. Because the candidates (i. e. ants) are influencing each other but are also eliminated after each iteration, this methods incorporates characteristics of both GA and particle swarm optimization [17], [32].

3.3.3 Embedded Methods

Embedded methods are hybrid models which incorporate FS into the training while still using the ML method to evaluate the prediction performance. They are more efficient as the computational effort is reduced if FS is included in the training due to smaller feature subsets. In addition to the model evaluation, a large number of features is penalized und thus faster excluded [48]. The advantage of embedded methods is reduced computational costs compared to wrapper methods [139] with SVR RFE as an example [108].

3.3.4 Peculiarities of Feature Selection

Various characteristics of FS are affected in the context of the present thesis and subsequently considered in more detail.

Introducing Artificial Features

Motivated from the previous section about relevance of features and instances, a further evaluation criterion for ML techniques is added which determines the quality of features and the value-add of these for the entire prediction. Adding enough irrelevant und artificial features differently distributed, any subset or only a feature for eventually just some instances will lead to a hypothesis with high predictive accuracy for at least a fraction of the validation set. Hence, any statistical induction method can explain anything if just enough irrelevant and differently distributed features are available in the Dataset (DS) producing many different feature subsets whereof all of these subsets or single features are used to learn a specific characteristic of the prediction target [83] also known as "If enough data is collected, anything may be proven by statistical methods" [41]. Using a no-information DS with only random variables, it is shown that already 100 artificially introduced features are enough to improve the accuracy significantly [83]. Also, the problem of overfitting is closely connected to features containing no information and is discussed in detail in the next section.

Hence, artificial features are created and added to the given DS and serve as comparison for irrelevant features to exclude the latter from the feature subset. Various approaches were investigated with Gaussian distributed [10] and risk assessed [153] artificial features as well as non-parametric variants [165]. A combination of the FOCUS filter technique and ID3 was shown to outperform the FRINGE and basic ID3 algorithms when artificial features were added for performance comparison [1]. Based on statistical analysis, the RELIEF algorithm was shown to be able to deal with noise and interactions between features [78].

Bias-Variance-Tradeoff, Overfitting & Underfitting

In ML a central aspect is the assessment of an induction technique by its results and the further development of the learning method depending on the outcome of the prediction. ROC curves or confusion matrices are popular criteria to evaluate classification methods, whereas the prediction error as MAE, RMSE (cf. section 3.6) or sum of squared error are common measures to investigate a trained model for regression [11], [175].

In order to ensure a generalized model for future predictions the problem of overfitting/ underfitting and the bias-variance-tradeoff need to be considered. Overfitting describes the effect of an induction technique to construct a function which does not learn the real, underlying pattern in the training DS. Instead, specific values in the training DS are extremely weighted or even memorized by the algorithm resulting in a falsified statistical model with excellent prediction accuracy on training data and poor prediction performance on validation or test data. Hence, a small training error due to an overfitted model and a high validation error are observable and occur in combination with small bias and high variance typical for flexible models which can be fitted very well to the given DS. A reduction of features, more instances within the training DS and increasing the regularization parameter are options to deal with this phenomenon. The induction algorithm might also almost be unable to infer the basic characteristics of the provided input and therefore it builds a generalized model which is too independent from the input DS (i. e. underfitting). Thus, low variance and high bias lead to similar predictions for all instances whereas a small difference exists between error on the training set and on the validation set because both are considerably high. Accordingly to overfitting, solutions to encounter high bias can be to include more features, to add polynomial features or to decrease the regularization parameter [52], [166]. Experimental results indicate that overfitting mainly occurs if the amount of available training instances is small. In addition to a validation set, a separate test set is holdout of the entire optimization process to estimate the real prediction performance of the final model. Although the validation set is used to validate the model during the learning process the algorithm might tend to fit the model to the validation set. Therefore a final test set with unseen data in the entire process is necessary to get a fair approximation of the prediction performance [83].

In a learning curve the total error can be drawn against the model complexity which is often coherent with a high number of features as shown in figure 3.2. During the optimization process to find the optimum model complexity and to avoid overfitting and underfitting, the total error has to be minimized to an optimum as indicated by the dashed line. The empirical risk decreases with increasing model complexity and a higher number of data available similar to the red line displaying the bias. Analogous to the total error the regularized risk increases again with increasing model complexity. Finally, the variance also increases with model complexity shown in blue.

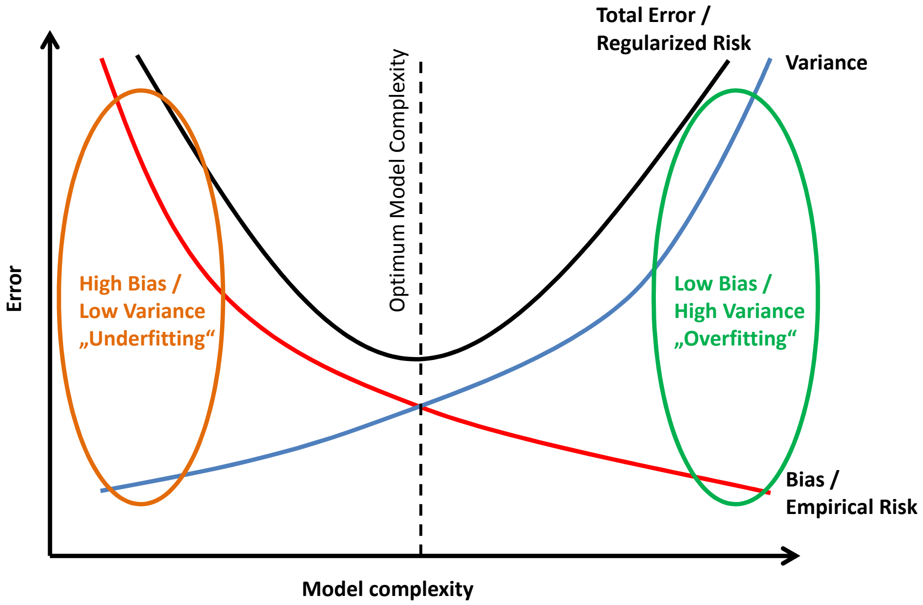


Figure 3.2: Bias-Variance-Tradeoff. Error is plotted against model complexity with an optimum in the center where the error is minimized and underfitting and overfitting are avoided [166].

Historical Feature Selection

An overview of FS methods is given in [145] with a chronological literature review and definition of the two basic approaches of forward selection and backward elimination. The first introduction of a wrapper algorithm combined with the claim to incorporate the induction algorithm into the process of FS was developed together with first definitions of relevance of features in 1994 [68]. The previously described problem of relevance of features and instances was initially discussed in detail in 1997 [12]. FS techniques are shown to improve the prediction performance in a variety of situations far beyond the aspect of presence of irrelevant features [12]. Comprehensive work was done to describe the FS wrapper approach for feature subset selection [81]. An early overview of FS algorithms as book [103] and the use of GA as search method in FS [177] were published in 1998. A correlation-based FS for ML was developed in a PhD thesis in 1999 [50]. In 2001 another book outlined approaches for FS [33]. Improvements by FS with SVM compared to other methods were shown in 2001 [173]. One year later, a book on subset selection in regression was written [115]. In 2003, a comprehensive summary and description of various features selection methods was composed with highest impact in the research of FS [48]. Subsequently, many enhancements were achieved in specific fields of research of FS with those related to GA and

SVM further highlighted in section 4.3.

3.4 Support Vector Regression

The classification with SVM as a supervised ML method can be extended to a linearized regression method for function estimation, i. e. SVR [151]. Applying a kernel function enables SVR to deal with nonlinear input data. SVR is based on Structural Risk Minimization [168] and finds a trade-off between model complexity and fitting the algorithm to provided data. The general idea is to search in the input dimensions for the most important vectors which best define a hyperplane between classes to be distinguished, the support vectors. A soft margin introduces a concept to deal with outliers. Using the Lagrangian, the saddle point yields the solution for optimization of the minimizing and maximizing dual problem [151].

3.4.1 Primal Optimization

In supervised learning theory we are given pairs of instances $(x_i, y_i, i \in \mathbb{N})$ from an input space $x \in X$ (e. g. $X = \mathbb{R}^d$) and an output space $y \in Y$. For ε-SVR (i. e. epsilon-insensitive-SVR) the aim is to estimate a function $f(x)$ with at most ε deviation from the target y_i for any input x_i [167].

Starting with the linear function for a separating hyperplane as given in equation (3.1),

$$f(x) = \langle w, x \rangle + b \text{ with } w \in X,\, b \in \mathbb{R} \tag{3.1}$$

the dot product between the weight matrix w and input data $\mathbf{x}$ as normal vectors is computed and added to the offset b based on a linear equation. To receive a flat function the margin $\frac{1}{\|w\|}$ around the hyperplane has to be maximized. So, the initial convex optimization problem is written as follows in equation (3.2) [151].

$$\begin{array}{ll} \text{M}\mathit{inimize} & \frac{1}{2}\|w\|^2 \\ \mathit{subject\ to} & \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases} \end{array} \tag{3.2}$$

To prevent overfitting, the ε-insensitive loss function defines an ε-tube (cf. figure 3.3-a) around the plane where all support vectors are considered to be equal [151]. The introduction of slack variables ξ_i, ξ_i^* as a soft margin concept (cf. figure 3.3-b) provides the opportunity to overcome the problem of otherwise infeasible constraints such as dealing with outliers. This is applied to SVM [9], [27] and is formulated in equation (3.3) [151].

$$\begin{array}{ll} \text{M}\mathit{inimize} & \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{l}(\xi_i + \xi_i^*) \\ \mathit{subject\ to} & \begin{cases} y_i - \langle w, x_i \rangle - b & \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i & \leq \varepsilon + \xi_i^* \\ \quad \xi_i, \xi_i^* & \geq 0 \end{cases} \end{array} \tag{3.3}$$

The cost factor C in equation (3.3) penalizes outliers for all l instances in $\mathbb{N}$ by setting the trade-off between tolerated deviations larger than ε (overfitting) and flatness of function f (generalization). Various types of loss functions exist. The used ε-insensitive loss function $|\xi|_\varepsilon$ is defined in equation (3.4) [52].

$$|\xi|_\varepsilon := \begin{cases} 0 & if \quad |\xi| \le \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \tag{3.4}$$

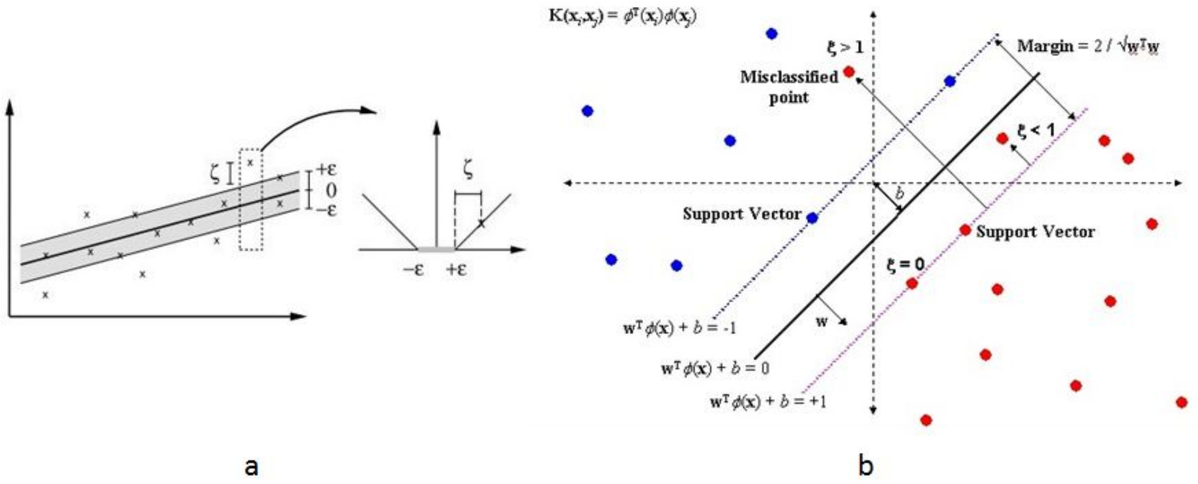


Figure 3.3: SVR Soft Margin: (a) ε-insensitive loss function [151], (b) Margin of ε-tube, support vectors and outliers [2].

3.4.2 Dual Optimization

The primal function f can be extended to a dual function by introducing the Lagrangian for the constraints. Thus, the saddle point gives the solution for optimization of the problem. The Lagrangian Multipliers $\eta_i, \eta_i^*, \alpha_i, \alpha_i^*$ are maximized while the primal function with respect to the weight w is minimized. The Lagrangian function L results as in equation (3.5) [151].

$$\begin{aligned} L := & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\ & - \sum_{i=1}^{l}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) \\ & - \sum_{i=1}^{l}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b) \end{aligned} \tag{3.5}$$

The Lagrangian Multipliers need to satisfy the positivity constraint in equation (3.6).

$$\eta_i^{(*)}, \alpha_i^{(*)} \ge 0 \tag{3.6}$$

The saddle point condition ensures the optimality and the partial derivatives of the Lagrangian function L with respect to the primal variables (w, b, ξ_i, ξ_i^*) equals to zero as given in equations (3.7)–(3.9) [151].

$$\delta_b L = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) = 0 \tag{3.7}$$

$$\delta_w L = w - \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) x_i = 0 \tag{3.8}$$

$$\delta_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0 \tag{3.9}$$

Now the equations (3.7)–(3.9) can be substituted into equation (3.5) to obtain the dual optimization problem in equation (3.10) [151].

$$\begin{array}{ll} \mathrm{M}axim\!ize & \left\{ -\frac{1}{2} \sum\limits_{i,j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle - \varepsilon \sum\limits_{i=1}^{l} (\alpha_i + \alpha_i^*) + \sum\limits_{i=1}^{l} y_i (\alpha_i - \alpha_i^*) \right. \\ subject\ to & \left\{ \begin{array}{ll} (i) & \sum\limits_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0 \\ (ii) & \alpha_i, \alpha_i^* \in [0, C] \end{array} \right. \end{array} \tag{3.10}$$

Substitution of w according to equation (3.8) into function (3.1) for a separating hyperplane yields the so-called support vector expansion (3.11) because w can be converted into a linear combination of x_i [151].

$$\begin{array}{rl} w & = \sum\limits_{i=1}^{l} (\alpha_i - \alpha_i^*) x_i \\ f(x) & = \sum\limits_{i=1}^{l} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b \end{array} \tag{3.11}$$

Hence, it is important to note that the complexity of a function represented by its support vectors does not depend on the input dimension X but only on the number of support vectors. Even in case of evaluating f, it is not necessary to calculate w directly. The entire algorithm can be computed as dot products between data motivating the usage of kernel functions to encounter nonlinear input data. The offset b of the hyperplane to the origin can be computed using the Karush-Kuhn-Tucker conditions. From these conditions it can be derived that for all input variables inside the ε-tube the Lagrange Multipliers α_i, α_i^* are zero and all vectors for which the coefficients (α_i, α_i^*) do not vanish are support vectors [151].

The hyperparameter C of the SVR model sets the boundary for all possible α_i, α_i^*, and thus defines how strong vectors outside the ε-tube are penalized.

3.4.3 Kernel Function

In order to allow SVR to deal with nonlinear input kernels are applied as a commonly used concept for mapping the input space X into a feature space F. Various kernel functions Φ :

$X \rightarrow F$ exist [167]. The kernel function k is only computed between the input data $\mathbf{x}$ as given in equation (3.12) [52].

$$\langle x_i, x_j \rangle \rightarrow \langle k(x_i), k(x_j) \rangle = k(x_i, x_j) \tag{3.12}$$

Thus, an important advantage of SVR is a nonlinear extension via kernels which does not influence the computing performance [151].

3.5 Genetic Algorithm

GAs, as introduced in subsection 3.3.2 as possible wrapper approach for FS, are evolutionary algorithms which derive their behavior from natural (biological) evolution. The search and optimization heuristic optimizes the solution using operations known from evolving population genetics. In general, an iterative GA creates a huge population of individual candidate solutions for many generations of the evolution to find the best individual which, in the present investigation, yields the best feature subset. Individuals are all spawned single possible solutions of the entire population whereas in contrast candidates are created within a generation as the possible parents for the next generation. Nevertheless, the terms individuals and candidates are often used interchangeably. For each generation, the best candidates from the previous generation are inherited to form an elite available for further evolution. From these elite candidates of the previous generation, new individuals are created in the current generation by the GA operations Mutation and Crossover thus reproducing these natural processes. Finally, these candidates are evaluated by a fitness function to find an optimum of the heuristic search and the best candidates are then inherited for the next generation again. The evolution of the GA starts with a chosen number of generations and the initial population of individuals can either be selected randomly or be seeded if a solution is expected in a specific range. As historically the genome of each individual is encoded as bit string with enabled or disabled genes and is altered and evaluated by the GA operations, nowadays the concept of a GA is mapped to a wide variety of optimization problems [117], [144].

Before listing an overall GA (cf. algorithm 1), the GA operations can be summarized as follows:

1. **Mutation**: Randomly flip bits on the encoding bit string representing the specific combination of characteristics of any individual (e. g. DNA).

2. **Crossover**: Flip all remaining bits to one end from a random location on the bit string.

3. **Selection**: Evaluate the initially defined fitness function for each candidate and select the best individuals.

4. **Inheritance**: Set the best individuals as parents for the next generation and inherit them to the next generation.

Algorithm 1: General sequence of a Genetic Algorithm

```
Data: Data for the problem to be optimized
Define Fitness function;
Create Chromosome as bit string from characteristics of population;
Create First generation from initial population;
while number of generations or threshold of fitness function is not reached do
    Crossover: Flip all genes to one end from a random location on the chromosome;
    Mutation: Flip random genes on the chromosome;
    Selection: Calculate the fitness function for each candidate and select the best
    individuals;
    Inheritance: Inherit from the best individuals as parents to the next generation;
end
Result: Best individual with regard to the fitness function
```

3.6 Evaluation Criteria

Many techniques exist to evaluate the outcome of ML algorithms whereas Recall-Precision, confusion matrix and ROC curves are typical for classification and Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are basic evaluation criteria for regression. For comparison reasons in terms of scale-independency the Coefficient of Variation of the RMSE (CV(RMSE)) is used in the present thesis. In addition to measure metrics like the precision, accuracy, specificity and sensitivity, it is expedient to estimate the goodness of fit or rather the explanatory power of the built model for which the Coefficient of Determination (R^2) is an appropriate measure.

3.6.1 Accuracy

The prediction performance in terms of accuracy is assessed by the following error measurements.

Mean Absolute Error

A basic error measurement is defined by the MAE where the absolute residuals $|\epsilon|$ as difference between predicted target values $\hat{y}$ and real target values y (i. e. $\epsilon := \hat{y} - y$) are summarized and scaled to n:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\epsilon| \tag{3.13}$$

Root Mean Squared Error

The performance of the used algorithms is evaluated by the RMSE. An advantage of RMSE as it implies a stronger focus on major deviations from the target compared to the MAE. Especially the ability of VM to detect crucial outliers should be increased to avoid further processing of affected (scrap) wafers. All residuals ϵ are at first squared, then summarized and divided by the total number of instances n. Finally, the RMSE is calculated as square root from the MSE:

$$RMSE = \sqrt{\frac{\sum\limits_{i=1}^{n} \epsilon^2}{n}} \tag{3.14}$$

Coefficient of Variation

The CV(RMSE) is introduced to provide a scale-independent value to enable overall comparability. The RMSE is normalized to the mean of all target values $\bar{y}$:

$$CV(RMSE) = \frac{RMSE}{\bar{y}} \tag{3.15}$$

3.6.2 Model Fit

The Coefficient of Determination (R^2) indicates how well the prediction model fits to observed data [16], [114]. It is a goodness of fit based on the empiric quadratic coefficient of correlation.

Thus, R^2 is most general defined as:

$$R^2 \equiv 1 - \frac{\sum\limits_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum\limits_{i=1}^{n} (y_i - \bar{y})^2} \tag{3.16}$$

Under certain conditions the empiric variance of the observed values equals the empiric sum of the predicted values and the residuals:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{3.17}$$

Holding these conditions lead to $0 \leq R^2 \leq 1$ in equation (3.18).

$$R^2 = \frac{\sum\limits_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum\limits_{i=1}^{n} (y_i - \bar{y})^2} \tag{3.18}$$

The higher the value of R^2 the higher the adaption of the trained model to the provided training dataset. For a value of 1 the model explains all the variance represented by data whereas a value of 0 means a prediction of the mean independent of any variable [37].

Some basic assumptions need to be considered in terms of applicability restrictions of R^2:

1. The coefficient of determination represents the goodness of a linear model fitted to data.

2. One the one hand a huge number of any available regressor, even without additional information, may cause a higher score of the R^2 whereas on the other hand a reduction of noisy and distracting regressors may also improve the result. Hence, to use R^2 as a meaningful evaluation criterion it should be aimed to significantly reduce the features to a minimal subset.

3. Finally, the target value is required to be the same for all compared models which indeed applies in the current approach [37].

High prediction accuracy in the process area of CVD was achieved by nonlinear ML methods (e. g. SVR with kernels) outperforming linear regression methods [128]. By fitting nonlinear functions to data also negative values for R^2 can be obtained [14] in contrast to equations (3.17) & (3.18).

3.6.3 Reliability

Sensitivity and specificity are common evaluation criteria to assess correct predictions for a binary test usually performed in classification. The sensitivity or so-called recall/true-positive-rate is defined in equation (3.19) as ratio of all correct predictions of positive test outcomes out of all positive test outcomes. In the scope of the present thesis and for regressions of continuous values, the sensitivity yields the rate of the correctly predicted outliers out of all outliers. In case of no available outliers, no value was assigned. Accordingly the specificity or so-called true-negative-rate is defined in equation (3.20) as ratio of all correct predictions of negative test outcomes out of all negative test outcomes referred as 'Non-Outliers'.

$$Sensitivity = \frac{\sum Cor.\,Pred.\,Outliers}{\sum All\,Outliers} = \frac{True\,Pos.\,(TP)}{True\,Pos.\,(TP) + False\,Neg.\,(FN)} \tag{3.19}$$

$$Specificity = \frac{\sum Cor.\,Pred.\,Non\text{-}Outliers}{\sum All\,Non\text{-}Outliers} = \frac{True\,Neg.\,(TN)}{False\,Pos.\,(FP) + True\,Neg.\,(TN)} \tag{3.20}$$

Summary: Following the required knowledge about SM and the complex HDP CVD process itself, the inevitable fundamentals of DM, ML, FS, SVR, GA and relevant evaluation criteria are embraced. The next chapter further specifies the introduced problems and presents a detailed review of the current state of the art emphasizing the enormous challenge to develop VM.

4 State of the Art

After introducing a detailed overview of Semiconductor Manufacturing processes and theoretical fundamentals essential for the development of smart Feature Selection to enable advanced Virtual Metrology, this chapter describes the state of the art of Feature Selection for Virtual Metrology, Virtual Metrology and Feature Selection themselves as well as prerequisites to enable advanced Virtual Metrology and smart Feature Selection.

4.1 Feature Selection for Virtual Metrology

The briefly stated challenges and problems (cf. section 1.1) for corporate-wide and advanced VM are derived from state of the art publications (cf. subsection 4.1.2) as well as previously performed work. These publications focusing on FS for VM are highlighted and discussed in order to emphasize the need for smart FS for advanced VM.

4.1.1 Challenges and Problems

The CRISP-DM approach (cf. section 3.1) is adapted to overcome the actual difficulties to consider and investigate only specific ML prediction methods and no further DP still yielding limited success. Since VM evolved as a major research area in SM during the last decade major effort is made to evaluate, enhance, adapt and optimize plenty of learning and regression techniques to finally achieve highly accurate predictions and to detect as many outliers as possible (cf. section 4.2). Nevertheless and in spite of partially quite successful implementations including reasonable schemes to deploy VM fab-wide, an advanced VM system solving all of the persistent problems could not be developed so far. As the challenges are mainly accepted by physicists, mathematicians and engineers with limited knowledge or experience in the area of DM and ML, so far no solution is found considering all the tasks and the full potential of DM and knowledge discovery. Thus, CRISP-DM is adapted to tap the full potential and to identify the remaining true challenges to implement corporate-wide VM.

Challenge 1: Business Understanding & Efficiency

The highly demanding competition in the SM industry requires to use the developed VM application in all suitable process areas and for the entire equipment variety as soon as possible in order to mandatorily maximize the return on invest (cf. appendix A.2). During CRISP-DM business understanding the entanglement of economic realization (cf. subsection 5.1.1) and technical challenges is elaborated. Regular iterative meetings including stakeholders and process experts are necessary to harmonize investment expectations, technical feasibilities and DM possibilities. This economical and technical challenge requires a generic VM approach which can be transferred and deployed without major additional effort. Hence, to achieve highest accuracy and reliability the approach of equipment specific FS and model optimization is required to run fully automated. While the adaption of statistical regression models (including model parameter optimization) is already available (cf. section 4.2), no FS techniques to corporate-wide automatically reveal only the crucial process parameters and to neglect noisy information has been presented so far (cf. subsection 4.1.2). Thus, the SM industry is still struggling to efficiently develop and implement automated fab-wide VM.

Challenge 2: Data Understanding & Scalability

In the area of logic circuit SM a variety of technologies, thousands of different basic types and numerous products are defined by the specific combination of available processes and the resulting multifaceted layer structure on the wafer. All these products are manufactured with many different operations and recipes which can be further subdivided into more specific logistical granularities. Finally, every wafer can be processed on a wide variety of manufacturing equipment and tools with some additional customized add-ons from various equipment manufacturers. In the context of data understanding as first iterative CRISP-DM core process, several topics related to data availability have to be considered. The huge manufacturing complexity emphasizes the enormous challenge to be able to handle data traffic, data storage and computational effort for many thousands of ML models for these logistical granularities all over the fab. In order to supervise, further enhance and develop VM, an obsolete prediction model still has to be kept and stored for several months up to some years since degradation of prediction performance and model behavior are mandatory to track. In case of deviating prediction performance in terms of accuracy or reliability, all actual models as well as predecessors need to be available for exhaustive inspection and drill down to find the root cause. All models have to be stored for deeper analysis as well as to reproduce any occurring error to fix possible bugs. Furthermore, the maintainability of a VM system of this scope with exhaustive usage of hundreds to thousands of equipment variables is infeasible since these equipment parameters can be constantly changed, recalculated, removed or new ones may be added by the process engineers. Thus, for the implementation and deployment of corporate-wide VM it is inevitable to reduce the amount of data storage, the accompanying massive data transaction, the immense computational work as well as the enormous maintenance effort. Hence, the development of a smart FS algorithm minimizing the number of input variables and concurrently maximizing the prediction performance in terms of accuracy and reliability appears to be imperatively necessary to overcome these obstacles and enabling the implementation of advanced VM.

Challenge 3: Data Preparation & Knowledge Discovery

The pure application and adaption of various ML techniques in fact can yield satisfactory results to detect outliers and to predict the real metrology with sufficient accuracy whereas time-consuming VM development and adaption processes are necessary including involvement of highly educated and expensive process experts from different manufacturing areas. Up to now, physical understanding is inevitable to select relevant input features for VM (cf. subsection 4.1.2, section 4.2). A manifold variation of truly illuminating and crucial, irrelevant and noisy as well as highly correlated input variables arises from underpinning data. Any selection of important features from these huge amount of possible input parameters by human expertise is usually done in a conservative way i.e. in case of doubt too much than too little features are extracted for subsequent VM modelling/application. VM developers, typically mathematicians or computer science experts dealing with statistical learning can either directly use these preselected features as input for adequate ML algorithms or prior to that may apply other dimensionality reduction techniques (e. g. Partial Least Squares (PLS)). After the time-consuming VM modeling process (not in a structured CRISP-DM manner but still including iterative feedback loops with process engineers), the prediction method of choice is applied to generate a VM model yielding the required prediction performance. Depending on the ML technique a trained VM model achieving good predictions is hardly or not at all interpretable by process experts (e. g. NN and SVR). Thus, on the one hand a process engineer needs to select input features to be incorporated into VM modeling whereas on the other hand the resulting decision which selected input feature subset is finally used is often not comprehensible to them because the trained VM models are based on complicated compositions and weightings calculated by the statistical algorithms (cf. subsection 4.1.2, section 4.2). Hence, a conclusion regarding the really crucial features out of the conservatively selected parameters can hardly be drawn and little additional knowledge is discovered to be used as basis for future process developments and improvements. A smart FS algorithm can provide an insight into the most important features really determining the complex physical process and thereby enable future process enhancements allowing the development of new products in logic device SM.

Challenge 4: Data Preparation & Accuracy

High product quality requires well-controlled manufacturing processes which in fact can only be achieved by very accurate physical measurements and thus by comparable VM predictions often tolerating not more than 1 % deviation of the real physical metrology as well as reliable outlier detection. Even though the application and adaption of ML methods already yield mostly satisfactory results, the prediction performance in terms of accuracy can be improved by FS during the task of DP in CRISP-DM to obtain remarkable results constantly less than 1 % deviation. In recently published work (cf. subsection 4.1.2) the potential of FS is recognized to reveal only crucial features containing valuable information and to discard noisy and redundant ones. According to the Bias-Variance-Tradeoff (cf. section 3.3.4), an optimization of model complexity is also investigated yielding the best trained model to be found at the minimum between underfitting and overfitting (cf. figure 3.2). Nevertheless, Principle Component Analysis (PCA), PLS and correlation-based reduction approaches are state of the art to achieve dimensionality reduction or rough FS (cf. section 4.3) in addition to expert selected feature subsets. As no high-sophisticated FS is invented up to now considering not only expert selected but all available features, the demand for corporate-wide universally applicable smart FS yielding highest possible accuracy increases.

4.1.2 Current Research

A Reliance Index (RI) between zero and one ($0 \leq \mathrm{RI} \leq 1$) is introduced in [24] to assess the reliability of a VM system. In more detail, the RI is calculated as the intersection between the statistical distribution of the applied prediction performed by a NN model and a reference Multiple Linear Regression (MLR) model. Furthermore, a similarity index is established to compare the actual instance for predicting the metrology outcome with instances from the historical training DS. If significant similarities occur a similar prediction is expected as well because the similar instance is included within the model training. Otherwise this specific prediction is handled more carefully [24].

NN for CVD: A VM scheme according to [102] for predicting CVD thickness in SM compares Radial Basis Function (RBF) NN and BPNN implementations for a single hidden layer and an adjustable number of neurons. The RI explained above is used to trigger real physical measurements if deterioration of the VM performance is indicated. Additionally, a module is integrated to automatically tune NN model parameters (e. g. number of neurons). Prior to any DP, an expert selection of 28 sensor variables is chosen for five important process steps resulting in 140 input features. The DP is broken down into five actions: 1) Missing values are populated with historical data, 2) latest data are used if the recorded time step exceeded five seconds, 3) mean and standard deviation is calculated for each variable, 4) data are normalized and 5) SS is performed to "pick comparatively important variables" [102]. Prior to step 5 and in order to achieve high processing efficiency, PCA is adopted to transfer input features into independent variables as required for RBF NNs. As it is common practice, data are split into training, validation and test DS. The relative MAE is chosen as evaluation criterion. The first drawn conclusion highlights the importance of DP whereas the second conclusion illustrates the superior performance of RBF NN compared to BPNN with high accuracy of ~0.4 % [102]. In spite of the high prediction accuracy combined with an elaborated VM scheme, the major challenge of an automated and generic FS approach applicable to other process areas with high efficiency could not be resolved. In order to achieve a more generic approach the fussiness of this scheme could be reduced by smart FS. Further interesting investigations according to [101] are conducted to perform FS by SS to predict an etch process in SM. Features are selected by a wrapper approach using SS with MLR and NN as learning algorithms as well as expert selected FS for comparison. Once more for the final prediction, BPNN is applied and superior performance is achieved by NN-based SS. The selection of 10 features out of 66 expert preselected features achieved a relative MAE of 0.89% and a maximum error of 1.7 % [101]. Nevertheless, due to preselected expert selected features neither algorithmic automatic FS nor corporate-wide efficient deployment is investigated.

NIPALS for CVD: Owing to the fact that methods like PCA, PLS and Nonlinear Iterative PArtial Least Squares (NIPALS) show limited prediction accuracy and ability to suppress noise components, an Iterative Backward Elimination IBE-PLS modeling technique is examined in [169]. The accuracy to predict the CVD film thickness measured by the RMSE and the ability to reduce noisy variables are compared to the NIPALS technique. The statistical models using initially 56 available features are trained and tested on 42 and 7 instances, respectively. The new approach of IBE-PLS demonstrated a considerable FS with 8 remaining features and improved the prediction accuracy to 0.75 % compared to 1 % deviation achieved by NIPALS [169]. This promising approach in fact demonstrated the potential of a good FS technique to enable the engineer to focus on the manageable remaining key variables. While the prediction accuracy yields superior results compared to PLS, the second challenge in terms of prediction performance to reliably detect outliers is not investigated. Also, the composition of a very small and specialized DS including only 42 training and 7 test instances with 56 features indicates quite restrictive preselection of features and instances and definitely illustrates the lack to tackle the problem of efficiency to develop and deploy a corporate-wide FS approach (cf. section 4.1.1).

Tree Ensemble for Etch: The application of ensembles of stochastic gradient tree boosting models to predict two levels of depth of a dry etch process in SM for different equipment chambers and different products is comprehensively investigated in [135], [136]. The embedded technique for FS is enhanced by a weighting of the categories of the predictor variables (logistical & equipment adjusted features vs. sensor features, mean & standard deviation values vs. compositions of several individual values). A so-called advanced method eliminating features showing more than 90 % correlation reduced the initial feature set from 120 to 80 features. The model is trained on a two months DS and constantly updated during the validation on remaining four months data. The mean and standard deviation of the residuals are used to examine the VM prediction quality. For the experiment the number of trained trees is set to 500 for a single gradient tree boosting model and the performance of ensembles of 20 and 65 models are evaluated in addition to mean & standard deviation in terms of calculated RMSE. Both ensemble collections achieved a prediction performance of 0.2 % and 0.8 % for mean and standard deviation of residuals, respectively. Also the Bias-Variance-Tradeoff is discussed within the scope of this work [135], [136]. The described VM approach applies a generic ML algorithm and introduces a FS technique which reduces the original feature set by 33 % by pure elimination of correlated features. Nevertheless, in terms of scalability for fab-wide VM deployment in other process areas with potentially up to thousands of features a reduction of roughly 33 % still too many features and even the permanent usage of 80 variables for the dry etch process generates high data traffic and substantial computational time for updating the model with every new available instance. The stunning prediction accuracy is neither comparable to other state of the art research evaluation approaches nor preferable since outlier detection is crucial and a basic requirement. The summation of signed instead of absolute residuals and a temporarily negative mean of residuals compared to the learning rate prevent a comparable assessment of prediction accuracy. Finally, the well-established approach of correlation-based reduction methods like filter approaches (cf. subsection 3.3.1) clearly demonstrates its limits to select only the most important features and to significantly reduce noisy variables. In the context of knowledge discovery and in contrast to smart FS tiny valuable information can be gained to enable process engineers to improve future developments.

PLS for PECVD: In preparation as input for future HDP CVD and Chemical-Mechanical-Polishing processes, a Plasma-Enhanced-CVD process for different layers, various products and two chucks within a single chamber is analyzed according to [40] as VM use case. The four main challenges accuracy, speed, throughput and flexibility considered in the publication corroborate the previously outlined challenges and problems (cf. subsection 4.1.1). 24 preselected FDC variables and 3 logistical parameters (i. e. chuck, layer, product) are incorporated to compare PLS consisting of four principle components with a tree ensemble method consisting of a combination of bagging from regression trees and random splits as main ensemble method. For 306 training and 168 test instances both ML techniques achieved a prediction accuracy of ~0.6 % RMSE and a coefficient of determination (i. e. R^2) of ~0.85 (cf. subsection 3.6.2). Out of the in total 27 features the tree ensemble approach selected five features whereof three are also included within the five most important features used by PLS. As conclusion, the PLS method performed slightly better and the introduction of a quality index to assess accuracy and reliability is motivated [40]. In spite of noticeable reduction of the feature set by ~80 %, a preselected initial feature set including only 24 process variables is investigated. Hence, this expert selected feature set already incorporates substantial expert knowledge and lacks the approval of being capable to significantly reduce bigger feature sets containing lots of redundant, noisy and correlated information. Apart from good accuracy, an assessment of reliable outlier detection is not conducted yet.

Canonical Analysis for PVD: In [147] a VM approach addresses the prediction of the resistance of glass substrates at 9 different locations in a Physical-Vapor-Deposition sputter process. On the one hand high dimensionality and collinearity in process variables are handled by dynamic canonical correlation analysis and canonical variate analysis whereas on the other hand expensive computational effort is encountered by VM model retraining after every new instance only if indicated by the RI thus optimizing frequent model updating (i. e. a MW approach). 20 variables are calculated from six sensor input parameters for 110 training instances covering four months productive data. The prediction performance yields an average of the relative MAE of 2.56 % [147]. The good accuracy demonstrates the need to consider the reduction of correlated features even though no sophisticated FS is performed to tackle the problems in terms of efficiency, scalability and knowledge discovery. However, the investigation of a RI as quality index as well as a MW approach to improve prediction performance is motivated. Finally, no outlier detection also crucial for corporate-wide implementation is considered so far.

SVR for Yield: A comprehensive and versatile inspection of SVR, PLS and rule ensemble as regression techniques and the two latter also as FS methods is conducted in [146]. A model is built based on 150 training instances and the MAE as percentage of the mean is calculated for 50 test instances. No real data are considered but the yield as prediction target is generated by two (linear & nonlinear) functions of all input variables which are generated by mathematical models. For the input variables, at first 10 key variables (i. e. mean of step 1 & step 2) are calculated out of 5 original FDC values, secondly in order to simulate correlated data each 20 linear and nonlinear variables as functions (i. e. linear, exponential & polynomial) of the key variables are created and at last 20 additional variables are generated as random noise using normal and uniform distributions. Both, the variable importance pareto as FS assessment and the prediction performance in terms of accuracy are evaluated within two major experiments containing various formats:

1. The first experiment investigates the target yield as linear combination based on the 10 calculated key variables with random Gaussian noise added. All methods achieved a comparable accuracy of 3 % to 4.5 % MAE with no major frontrunner. Rule ensemble performed superior compared to PLS in terms of selecting the most important features because it really revealed the most significant feature while PLS could not achieve a strong differentiation between this feature and another linearly generated one. However, both methods captured the actual top three contributing features. Subsequently, key variable drifts and shifts are introduced resulting in a degradation of the prediction performance from 4.4 % to 5.9 % and 4.7 % to 6.1 %, respectively, whereas rule ensemble only achieved an inferior MAE of 10 % for key variable shifts. Hence, in the linear setup PLS and SVR performed best with respect to prediction accuracy with a slight advantage for PLS due to the easier model interpretability. In contrast, rule ensemble outperformed PLS to reveal the most significant feature.
2. The second experiment examines the target yield as nonlinear function with input variable correlations and random Gaussian noise. The initial evaluation of nonlinear simulated features yielded a strongly degraded prediction performance of 11 % to 17 % with rule ensemble performing slightly better with 11 % to 14 % deviation from the target. Introducing key variable drifts and shifts results for PLS and SVR in almost unchanged prediction performance between 14 % and 19 % whereas for the rule ensemble technique the performance drastically dropped to 28 % and 23 % for included drifts and shifts, respectively. Finally, also the assessment of the actual variable contribution demonstrated a more differentiated variable importance pareto for rule ensemble compared to PLS. Both methods captured several top contributing features but due to the poor prediction performance of rule ensemble in the nonlinear setup its FS outcome can be assumed to be inferior.

Nevertheless, the conclusion points out that the usage of rule ensemble is not recommended in the linear case due to an inferior prediction accuracy and comparable FS performance. SVR slightly outperforms PLS in terms of prediction accuracy and both also outperform rule ensemble with remarkable differentiation whereas the latter is stated to provide significantly better variable contribution results than PLS. While apparently no technique provides reasonable prediction accuracy with nonlinear yield models more analysis and improvement is demanded [146]. It is hard to compare the results of the three techniques since every time different test sets are used for evaluation. In spite of the concluded comparable FS performance in the linear case a clearly visible difference in revealing the most important feature comes out. Even though SVR and PLS are stated to be comparable in the nonlinear case SVR is suggested for further investigations since its full potential is unlikely to be exploited. Regarding the lack of direct FS ability of SVR, a FS SVR wrapper approach could serve well to tackle the problem of accuracy within FS due to the fact that the prediction performance turned out to be deficient in this investigation with at best 3%. The extent to comply with the challenges of knowledge discovery and scalability can only be assessed with limited precision since only a few of the real variables are listed in the variable importance pareto and the yield prediction lacks in accuracy. Regarding the challenge to master the demand of efficiency and corporate-wide VM deployment, the capability to achieve a prediction performance of $< 1\,\%$ for both the linear and nonlinear case appears to be mandatory. In conclusion, potential and need of smart FS generically enabling advanced VM prediction is again corroborated by the results of this investigation.

Recursive Coefficient Centering for Critical Dimension: In a memory fab in the area of plasma etch the selection of important plasma sensors variables is analyzed to enable robust VM covering shifts across preventive maintenance cycles through a cost-effective recursive coefficient centering technique (i. e. data normalization) [6]. The FS technique describes an integration of squared sensor variables ranking to perform an interaction analysis which basically equals a time-integrated variance calculation method. MLR and PLS are used as linear regression methods to predict the critical dimension (i. e. another prediction target related to the common etch depth). 7 expert selected FDC variables (e. g. pressure, bias power, gas flow) are combined as input feature set in a 18 x 7 matrix with 18 sensor variables which are also selected by process experts from optical emission spectroscopy. For 25 training and 45 test instances an adaptive VM model is examined and updated in coincidence with a performed wet clean maintenance after the 10^{th} test instance which is referred to as "when VM lifetime ends" [6]. Achieving a relative MAE of 3 % and higher than 7 % after the periodical maintenance, the conclusion in [6] is drawn that linear regression methods are useful for the VM approach even though outlier detection is excluded due to the average evaluation result of 5 % accuracy. With respect to efficiency and scalability no assessment can be obtained due to the fact that the investigation was specific for a single use case and the feature set is statistically calculated from a 18 x 7 matrix. The restriction to exclusive expert selection of the input variables precludes a FS process to discover further knowledge. In general the approach examines a new statistical convolution methodology for an aggregated prediction target whereof the accuracy assessed by the relative MAE can hardly be compared to current research using direct prediction targets like etch depth or deposition thickness.

GA FS Wrapper for Etch: An application of FS for VM is conducted according to [179] in 2008 for the plasma etching process and focuses on DP in terms of data integrity and quality to obtain a robust prediction model. Outliers are removed by PCA and 3 types of FS techniques are compared for PLS, Stepwise and BPNN as VM models: i) Random modeling, ii) GA search as FS wrapper and iii) SS & clustering. Based on 500 training and 270 validation instances and with an initial feature set of 149 variables, the best algorithm (i. e. BPNN with SS & Clustering) yields 37 features and R^2 of 0.74 as "fairly good prediction results" in [179]. This work is extended in [178] with suggestions of various possible techniques whereas data separation based on PCA and mean replacement for missing data is chosen. Outlier detection is performed by investigating Hotelling T^2 (cf. [172]), by PCA on input covariance data in order to compute the distance and by statistical distance plots derived from PCA. Updating actual training instances as MW approach are selected by PCA based on similarity factors and clustering. As last preprocessing method, FS is again conducted by regression coefficients, a GA search with PLS and SS. Finally, three regression methods (i. e. BPNN, PLS & stepwise regression) are used to predict the target of the etch process with again "fairly good prediction results" (i. e. RMSE of ~0.5 %) [178]. In this research, a preceding clustering technique using logistical parameters is examined in difference to individual categorization or model configuration. The opportunity to dynamically configure many dedicated and specific VM models in order to create a generic VM system to achieve high efficiency for corporate-wide deployment is not addressed so far. Noticeable is the dimensionality reduction of roughly 75 % of all initial features and the very accurate prediction performance. The applied outlier exclusion by PCA points out the missing focus on outlier detection which is crucial to detect scrap wafers and thus to achieve the required highest level of product quality in SM.

Clustering for Etch: According to [126] in the SM area of plasma etch, the development of robust, reliable and interpretable VM models is a big challenge due to the highly correlated input space of available data. Maximal separation clustering as an unsupervised correlation-based clustering algorithm is applied as preprocessing step to identify input variables for the four FS techniques forward selection regression, forward selection ridge regression, ridge regression and LASSO (Least Absolute Shrinkage and Selection Operator). For 2000 instances the normalized MSE is calculated to evaluate the best input features made up from four statistical moments (i. e. mean, variance, skewness & kurtosis) of the amplitude of the time series for each of 2000 wavelengths recorded by optical emission spectroscopy. Training, validation and test sets are split into 50 %, 25 % and 25 % of the available DS, respectively. The achieved improvement of the prediction accuracy by the maximal separation clustering method of 13 % for forward selection ridge regression and 8.5 % for forward selection regression are contrary to the stated degradation for ridge regression and LASSO. A maximum increased prediction accuracy from 3.74 % down to 3.2 % is obtained [126]. The given publication deals with the challenge to reduce a massive amount of interrelated input parameters using a correlation-based unsupervised clustering method. In spite of little absolute improvements of the prediction accuracy, no significant optimization can be achieved using the unsupervised approach as preprocessing step thus still not achieving highest required accuracy of less than 1 % deviation from the target. Similar to a preceding statistical filter technique the massive amount of input features (i. e. 8000) is significantly reduced to 131, 270, 979 or 4888 features left with respect to the FS technique. Nevertheless, in order to attain scalability the number of features is still too high and also the goal of knowledge discovery can hardly be achieved for this amount of features to be analyzed.

SVM for Outlier Detection: Other VM ML-based research according to [76] focuses on novelty detection to reveal faulty wafers. This approach is motivated by the lack of unique characteristics of outliers within an imbalanced DS. Similar to the previous methods, Stepwise Linear Regression, PCA and SVM are applied as FS techniques where 150 features are present as real FDC data. Seven different novelty detection methods are tested. Not the final reduction of the initial feature set but sensitivity and specificity are the targets to be evaluated differentiating between two experimental settings: Cross-Validation and MW. In spite of an overall good True-Positive-Rate (i. e. sensitivity/outlier detection) of ~ 78 % is achieved, the poor False-Positive-Rate (i. e. specificity/"normal wafer classification within limits" [76]) of ~ 40 % motivates for future research to control the True-Positive-Rate vs. False-Positive-Rate trade-off by determining misclassification costs. It is outlined that cross-validation outperforms the MW approach which might be caused by a too small training set for MW and thus is also part of future work [76]. No concrete evaluation of the FS methods in order to reveal only the most important features and their impact on VM was performed. Thus, the challenges outlined in subsection 4.1.1 are not considered and cannot be assessed in more detail.

Aggregative Linear Regression for Etch: Another VM application in the area of plasma etch compares Forward Stepwise Regression, DTrees and correlation techniques for dimensionality reduction and also investigates a systematic approach to evaluate confidence intervals generated by Aggregative Linear Regression, MLR and Gaussian Process Regression [125]. 1894 instances are used for training and 300 to test the obtained ML model. A huge number of 12288 input variables is made up by 2048 wavelengths from an optical emission spectroscope and 6 statistical moments (mean, variance, skewness, kurtosis, max, min). Forward Stepwise Regression as best method outperformed DTrees and correlation techniques in terms of accuracy (i. e. normalized MSE 2.2 %) and dimensionality reduction (i. e. 346 features left). The smallest and best confidence interval could be achieved by Aggregative Linear Regression compared to Gaussian Process Regression and MLR [125]. The accuracy is in general not unconditionally sufficient for highest requirements for VM in SM while the drastically reduced feature set (9700 eliminated variables) is impressive but also expectable for an input of more than 10.000 features. However, 300 input features are still far too much and thus the problems of efficiency, scalability and knowledge discovery cannot be tackled successfully.

FS and Projection for Etch: Considerable research in the field of FS and VM is initially performed in 2009 [70] investigating two subsequent etch processes (P_1 & P_2) over a time period of three month. Two objectives are defined: assessment of the prediction specificity $(TN/(TN+$

$FP)$) (cf. equation (3.20)) which is the model accuracy in terms of correctly classified wafers within the control limits and the sensitivity ($TP/(TP + FN)$) (cf. equation (3.19)) which is the ratio of outliers detected by the VM system compared to undetected outliers. As it is good practice, all data are normalized to [0,1]. The total number of 1546/1793 input features of process P_1 & P_2 is composed of four key numbers (i. e. min, max, mean & stdev), each for eight process steps with each step containing 48 features (i. e. equipment sensor variables) for P_1 and 56 for P_2. These 1536/1792 predictor variables are completed by 10/1 metrology variables resulting in 1546/1793 input features. Here, 118 and 241 instances are available for P_1 & P_2, respectively. Three targets $T_1 - T_3$ are defined for P_1 and a fourth target T_4 for P_2 resulting in four tables. Finally, the DSs of each target T_i are subdivided into: 1) an overall DS with all selected equipment sensor variables and all metrology input features and 2) a second selected DS also containing all equipment sensor variables but only one additional metrology input feature. Thus, four tables for all targets T_4 are generated with each an overall and selected DS. Two FS methods and two feature projection methods (cf. section 4.3) are tested for each target T_i: FS as Stepwise Linear Regression as well as GA-based FS with SVR as induction technique and PCA and kernel PCA as projection algorithms. These four FS/projection methods are each tested for the overall and selected DS resulting in eight rows within each of the four tables of target T_i. As columns, five regression methods are subsequently used for final prediction: Linear Regression, k-Nearest Neighbor, NN, Regression Trees and SVR. Finally, four FS/projection techniques and five prediction methods are compared each on two DS (i. e. overall & selected) for four targets $T_1 - T_4$ (i. e. $4 * 5 * 2 = 40$ entries in 4 tables). In general over all DSs, the Stepwise Linear Regression FS algorithm performs best in terms of accuracy with Linear Regression, NN and SVR as prediction methods. A remarkable reduction of 84 % (i. e. 1307 eliminated features) up to 99 % (i. e. 1538 eliminated features) of all initial features is achieved yielding a prediction performance in the range of 0.53 % up to 2.54 % with the best result of 0.53 % accuracy by elimination of 99 % of all features. In the end, all wafers within the control limits are correctly classified (specificity = 100 %) and the majority of the present outliers is detected (sensitivity = 65 %) [70]. The result of 100 % specificity is less insightful since the very broad control limits cover a range of 0.97 up to 1.06. However, the obtained and quite good sensitivity of 65 % corroborates the possibly mastery of the challenge to detect most outliers. A first limitation stated in the publication covers the missing investigation of process drifts due to the small number of 118 and 241 instances for P_1 and P_2, respectively. Secondly, the problem of fab-wide VM implementation to enable efficient deployment (cf. section 4.1.1) is not tackled so far. The noticeable feature reduction of up to 99 % demonstrates possibly achievable scalability and knowledge discovery. Finally, the accuracy in terms of MSE from 0.53 % up to 2.54 % is partially remarkable but again in general not unconditionally sufficient for highest requirements for VM in SM.

Summary FS for VM: In general, many quite simple but also very sophisticated FS approaches for VM have been developed. This section provides a comprehensive literature survey regarding FS for VM and highlights the increasing demand to find a solution to tackle all problems stated in subsection 4.1.1. Various FS methods are investigated for several process areas but the challenges of concurrent efficiency, scalability, knowledge discovery and accuracy could not be mastered so far.

4.2 Virtual Metrology

Following FS and projection for etch, further work [69] in the process area of lithography suggests to enhance the previous concept by improved performance of a subsequent but simulated R2R controller which also manages to better deal with outliers. If no process drift is observable a longer VM training period is preferred for better prediction accuracy whereas in case of steady process drift a MW approach is motivated [69]. The evaluation of the significance of VM input for R2R control emphasizes once more the requirement for an advanced VM system to improve process control in SM.

For successful application of VM reliable and accurate historical data are crucial to train any ML model [58]. The quality of available data varies significantly between different process areas and work centers according to data collection within different systems, different production and metrology equipment as well as interfaces used for data acquisition, conversion and preprocessing. In most cases, missing data, outliers or fragmented data appears to be inevitable. So, in terms of Knowledge Discovery and DM, DP has become most essential to obtain a purified data set for successful DM [103], [129]. DP comprises several tasks (e. g. outlier removal, missing data deletion/conversion, etc.). In order to obtain accurate, reliable and reproducible VM results, comparable high effort has to be effected for DP. As the trained VM model needs to deliver reliable prediction of the metrology outcomes, the used ML algorithm has to be robust towards any shifts and drifts of the input parameters whether they are related to preventive maintenance actions (e. g. process chamber wet cleans) or intrinsic changes of process conditions (e. g. due to process chamber deterioration or contamination). Also important for the success of the implementation of VM is the iterative involvement of process experts for a priori data analysis as well as selection of relevant parameters in addition to the application of FS algorithms in order to keep the computational effort within feasible limits. Regarding the appropriate choice of prediction algorithms, it already becomes obvious that Simple Linear Regression and MLR are not unconditionally suitable for VM in SM due to their lack of robustness and accuracy [128], [155]. Further research in this area focuses on robust and high-sophisticated statistical models (e. g. Classification and Regression Trees [40], Neural Networks [155]). Recently, SVR evolved as a new promising state of the art regression method feasible for accurate and reliable VM modeling.

VM at TSMC: An early VM system was implemented at TSMC foundry in 2005 for shallow trench isolation deposition and plasma etch [21]. In order to keep pace with Moore's law, shrinking device dimensions are necessary and with it Wafer-to-Wafer control evolves as critical requirement. While it is economically infeasible to realize 100 % physical metrology for every wafer, novel methodologies, in particular VM, are inevitable to assure process control and quality for each wafer. VM enhances APC by improving the so-called centered process capability index [62]. The deployed VM implementation at TSMC achieved a coefficient of determination (cf. subsection 3.6.2) of at least 0.97 for the shallow trench isolation CVD process and higher than 0.98 for the associated etch process [21]. Thus a significant improvement of the centered process capability index was achieved by deployment of the VM system. Nevertheless, due to no performed FS neither efficiency nor scalability (cf. subsection 4.1.1) are considered and so corporate-wide deployment is not recommended if in future high performance of VM systems is required.

Cheng et al.: In further work, a conjecture NN model predicts the target process outcome and a second prediction model using weighted moving averages incorporates the obtained real metrology as well as the quality and the predicted accuracy of the first model for self-adjustment [154]. Considerable improvement to the previous conjecture NN model and noticeable impact is caused by the contribution of [25] to develop a Dual-Phase VM scheme. The first phase focuses on a fast VM prediction by a NN algorithm for a CVD process in a semiconductor TFT-LCD monitor fab including the computation of RI and a similarity index (cf. subsection 4.1.2). The second phase collects the metrology of sampled wafers and correlates it with the stored prediction. In case of major deviation the VM models (i. e. NN) including the RI and similarity index values are retrained and subsequently updated. Finally, the VM prediction is recomputed for the entire lot. Again, the relative MAE is evaluated as threshold for possible retraining. A relative MAE of 0.6 % to maximum 1.2 % confirm the correct approach to encounter the problems in VM [25]. According to [25] the prediction accuracy of the developed VM scheme ranging from 0.6 % to maximum 1.2 % relative MAE is achieved without focusing on outliers whereas especially these outliers are most interesting since they indicate a process failure and possible product breakdown. Thus, using RMSE as evaluation criterion as well as including the occurred outliers into the tested DS will provide a more meaningful result with also little degradation of prediction performance in terms of accuracy. The same subject of missing but essential outlier focusing appears in the publication of [101] which already addresses the objective to FS for VM. Even though after expert selection SS was examined demonstrating a prediction performance of 0.89 % up to 1.7 % relative MAE but again without considering outliers within any tested data. Various other extensions of the activities of Cheng at al. are published with focus on NN algorithms and application for CVD processes for TFT-LCD manufacturing. In this context a generic VM framework with application drivers and interfaces is designed and induction methods (i. e. NN) as well as a RI and a similarity index are implemented [57]. A strategy to perform fab-wide VM deployment distributed on various VM systems is proposed by this research group in 2008. A central model-creation server generates and fans out the induction models to the distributed automated VM servers. The proposed strategy is verified including an automatic model retraining approach. A feasible and affordable deployment is achieved by reduction of fab-wide manual model creation [59]. VM prediction with ML methods (i. e. NN and MLR) are tested to meet real-time requirements and provide acceptable accuracy showing similar results as in [128] where MLR failed in terms of accurate predictions. BPNN with two hidden layers is too time consuming due to the chosen setup with an exponential influence of each hidden layer and thus this approach is regarding the computational time effort inferior in productive application compared to single layer BPNN [155]. Recently, the widely common SM execution system is extended by this automated VM system according to [23] where a model creation server generates and fans out new VM models. So, Database (DB) and VM client-server solutions are coupled with R2R control. The projected goals of cycle time and cost reduction of a modern SM fab are accomplished [23], [60]. In order to integrate VM into the entire SM fab environment a solution is presented in [71] to assign a weighted factor depending on the RI to the VM prediction as enhanced input to the R2R feed backward loop. Here, the originally calculated R2R coefficient which is based on real metrology is multiplied with the RI which itself is rescaled between 0 and 1. Adding the RI improves the centered process capability index compared to integrated VM without RI [71]. In [176] a dynamic MW scheme deals with the challenge of keeping the training DS in productive environment up-to-date to ensure high prediction accuracy for recently built VM models. Using ML similarity clustering the training instances are categorized. A limited cluster size and the 'First-In-First-Out' principle ensure the dynamic MW DS to stay up-to-date [176]. An automated VM system with model-creation server and connection to R2R control as given in [23] serves as an example for the indispensable requirement of an integrated VM system in order to efficiently assess the feasibility and prediction performance.

NN for Chemical Mechanical Planarization: For VM in SM also a combination of piecewise linear NN and fuzzy NN is investigated where these algorithms are applied to predict the target including prediction of process drifts by the former and the influence of process recipes (i. e. process shifts) by the latter. The VM approach is applied for a chemical mechanical planarization process using MSE to measure a prediction performance of ~ 0.75 %. Hence, sufficiently high accuracy and generalization ability even regarding recipe adjustments are proven [19]. As the good results are achieved for the provided use case without any FS performed no progress is made in order to develop an efficient and scalable VM system including the ability of knowledge discovery.

Wafer-fine R2R Control: Comprehensive work with significant impact was published in 2008 in [75] describing a strategy to connect VM and R2R control to implement a fab-wide Wafer-to-Wafer control system. Both, feed forward and feed backward control are outlined. A feed forward approach is realized by providing the VM prediction as a further source of information in addition to the real metrology for the R2R controller of the subsequent process to enable wafer-fine R2R control. A feed backward approach is designed by returning the VM prediction to the R2R controller which just calculated the adjustments of the current process. Hence, a retuning of this R2R controller can be done immediately. FDC and metrology data serve as basic data source. PLS Regression is applied as linear multivariate method for prediction and simultaneous FS of the given multiple-in-multiple-out process. Input variables preselected by experts and the initially from design of experiment obtained regression model are periodically updated by a recursive MW approach which noticeable improved the prediction performance compared to a fixed design of experiment based model. The MSE and the prediction output variance as ratio are used to monitor the R2R controller. An improvement of wafer quality compared to Lot-to-Lot control is observed. Furthermore, the effect of delay of physical metrology on R2R control is eliminated by the wafer-to-wafer control system. Finally, without concrete and quantifiable results the investigated feed backward control including VM achieved satisfactory performance [75]. A similar approach focusing on two consecutive processes (i. e. lithography, plasma etch) is published by the same authors half a year earlier also using PLS as regression technique illustrating that a feature preselection by process experts does not solve the problems of efficiency, scalability or knowledge discovery [74]. Unfortunately, missing quantified assessment of the result hinders the comparison of the introduced recursive PLS approach to other work. Nevertheless, in the outlook of the paper more research to find a fab-wide control strategy to implement VM is demanded.

4.2.1 Virtual Metrology in IMPROVE

The 2009 - 2012 performed European research project "Implementing Manufacturing science solutions to increase equiPment pROductiVity and fab pErformance (IMPROVE)" [63] was initiated by ENIAC (European Nanoelectronitcs Initiative Advisory Council) to improve the state of the art research and development in European SM with focus on APC. Within the project, various clusters of semiconductor manufacturers, research institutes, universities and solution providers were formed to focus on the three most important future APC areas including VM as one of them. Lots of research and productive implementation is conducted yielding about 100 internationally acknowledged publications whereof the following are dedicated to research and application in the area of VM.

DTree for CVD: Initial investigations to predict the CVD layer thickness based on FDC data are focusing on two variations of MLR and three variations of NN which perform superior in terms of prediction accuracy measured by the MAE [39]. Further work in collaboration with AustriaMicroSystems addresses the challenge to simultaneously meet accuracy, speed, throughput and flexibility requirements in VM development. The PLS method achieves better prediction performance in terms of RMSE than the compared DTree [40]. While no FS is investigated the assessment still indicates the effort to optimize prediction performance only by varying ML techniques.

Forward Selection Component Analysis for Etch: FS (i .e. Forward Selection Component Analysis & Forward Selection Regression) and feature projection methods (i. e. PLS & Principle Component Regression) are assessed with highly correlated data for a plasma etch process yielding superior prediction performance for the FS technique compared to projection methods in terms of a smaller error (normalized MSE) due to real feature reduction in contrast to feature weighting done by the projection technique [131]. Afterwards for the same setup, PCA is compared to SS with either MLR or NN as regression algorithm and RMSE as evaluation criterion. Also, a disaggregation method is introduced which splits the DSs into three parts still chronologically ordered and related to the maintenance cycle. On the full DS PCA with NN produced the best result (i. e. 0.99 % RMSE) whereas on the disaggregated DS the combination of SS and NN performed best (i. e. 1.18 % RMSE) [107]. Additional work is performed on disaggregation where a chronologically ordered and interleaved global DS and a separated, clustered and windowed local DS are assessed with PLS, NN and Gaussian Process algorithms. Finally, it is shown that a local MW DS with Gaussian Process regression yields the highest accuracy with relative MAE = 1.14 % [106]. Further investigations on Gaussian Processes reveal an accurate prediction which is rather insensitive to the used covariance function. Due to the fact that the prediction is obtained as distribution, engineers can quickly create confidence intervals to estimate the possible degree of variation for each predicted value [105]. The comparison between FS and projection methods corroborates the preferable approach to tackle the problem of high dimensionality by FS because efficient knowledge discovery for further process development is hardly possible by feature projection methods. However, even with an accuracy ~1 % the challenges of efficiency and scalability are not investigated.

$\mathcal{L}_1$-penalized ML for CVD: At Micron Technology SM fab, a hierarchical framework based on a $\mathcal{L}_1$-penalized ML technique is developed for a CVD process. Particularly, the LASSO (Least Absolute Shrinkage and Selection Operator) method is successfully extended to a more generic multi-level LASSO algorithm enabling VM predictions on nested levels of variability [122]. Even though the hierarchical framework is designed to encounter the problem of efficient corporate-wide VM deployment, no concrete results regarding any of the four stated challenges are presented for a more detailed comparison.

Software Framework: In order to encapsulate the development of smart ML prediction techniques as well as to enable plug & play integration, a generic SM fab framework is developed within the scope of the IMPROVE project [142], [134]. VM and Predictive Maintenance are both applications relying on historical productive data and aim to predict either the metrology outcome or the time when the next maintenance needs to be scheduled. Within complex SM, a wide range of IT systems is available for specialized applications including commercial as well as in-house developed APC software systems with many individual DBs and interfaces. Thus , an appropriate state of the art solution for efficient fab integration is designed, implemented and deployed for pilot testing in the course of the project. The framework services are implemented as Enterprise Java Beans deployed in a JBoss Application Server environment with CORBA procedures enabling a company specific development of VM algorithms either in MATLAB or R programming language as well as a high level of genericity and a remarkable good performance [142], [134].

4.2.2 Virtual Metrology at Infineon

Three Infineon frontend manufacturing sites (i. e. Regensburg, Villach, Dresden) were involved in the IMPROVE project. Various approaches have been tested during VM development resulting in the VM system described in chapter 5.

In [156] popular PLS and BPNN techniques are also investigated at Infineon Austria with frontend manufacturing site in Villach for a CVD process based on FDC data. Data clustering is identified to be critical in high-mixture SM environments if insufficient data are available to represent the logistical variation (e. g. various operations, technologies, products, recipes). In case of constant manufacturing conditions and low variation the more accurate BPNN method is recommended whereas PLS should be considered otherwise [156]. On the one hand the stated "satisfactory" [156] prediction performance of 0.45 % MSE for BPNN and 0.92 % MSE for PLS might be improved and on the other hand no real progress regarding to the four key challenges (cf. subsection 4.1.1) is advanced made as FS is not in scope of the investigation.

The already discussed multilevel LASSO algorithm [122] is developed in collaboration with Infineon Austria and successfully deployed and tested within productive SM fab environment but without comparable evaluation metrics [143]. Statistical inference is not yet present within VM but first ideas have been implemented and tested as multistep VM approach for SM processes. For CVD, thermal oxidation, coating and lithography as four subsequent manufacturing processes, three prediction scenarios are compared including more or less information in terms of sensor data and logistical information. The most complete scenario considering process data of CVD and lithography as well as logistical information of all four process steps shows superior prediction performance. Sample size and relevance are identified to be important and the multistep VM approach in fact reduces the calculated RMSE vs. single step VM but again without comparable evaluation metrics [123]. Furthermore, an interesting simulation on integration of VM into R2R control is assessed based on information theory. The Kullback-Leibler divergence is applied to evaluate the prediction quality and the Shannon entropy is used to reduce the risk of drifting VM predictions vs. real metrology sampling. Tests compared to the RI approach show superior performance. This aspect is closely connected to the MW approach to find the best historical DS on which an actual VM model is trained. Finally, the investigated methodology is concluded to outperform actual R2R control with correlation-based noise evaluation and poor VM prediction [158].

In terms of a review of regression methods for prediction of deposited layer thickness for a CVD process, an early Infineon publication [128] is already referred to in previous sections. First considerations on DM and ML techniques are published in 2011 according to [92] with focus on SVR. A PECVD process is chosen and a first assessment is performed to identify which prediction target yields the most accurate prediction in dependence of various logistical granularity. As a result, the prediction of the deposition rate achieves better performance compared to the prediction of the deposited layer thickness itself [92]. The IMPROVE framework, as briefly described above, is also tested at Infineon during the project [91]. A comprehensive review of the following regression methods is conducted for VM in the CVD process area: Simple Linear Regression, MLR, Ridge Linear Regression, PLS Regression and SVR. The infeasibility of Simple Linear Regression and MLR are proven. PLS and Ridge Linear Regression clearly outperform MLR but due to the linearity assumptions, these methods are outperformed by SVR which shows the best generalization on test data [127]. Further research on SVR-based ML to enable highly accurate and reliable VM at Infineon is corroborated by remarkable results. A CV(RMSE) smaller than 1 % and the equivalent CV(MSE) equal to 0.7 % approves high predictive power of SVR as well as the robustness of the model which is also indicated by a R^2 of 0.64 % [90]. Focused on ML research, various successfully implemented high-sophisticated induction techniques (i. e. DTrees, NNs & SVR) based on different statistical and mathematical theories including different assumptions and boundary conditions are compared. The results demonstrate the robust and highly accurate prediction performance to enable reliable VM being able to detect outliers in the target values and thus misprocessed wafers. Even on small sized DSs common in low-volume-high-mixture SM production the investigated algorithms perform superior to others [127] with an overall relative deviation smaller than 0.5 % on independent, unseen test data which appears to be close to the achievable minimum with regard to the typical accuracy of physical metrology. Comparable accuracy of these methods enables the usage of the quantified similarity of the prediction results as new index to measure the reliability of a VM prediction. In case of major deviation between DTrees, NNs and SVR, the VM outcome should be treated with special caution and real metrology sampling appears to be recommended [93].

Summary VM: Only few VM schemes demonstrate the opportunity to develop generic VM systems appropriate for corporate-wide deployment and economically efficient implementation of VM in SM. Several publications consider the possible enormous amount of up to more than 10.000 input features and provide solutions to encounter the so far not clearly formulated challenge of scalability required for corporate-wide VM deployment. Some research has been performed regarding FS for VM to reveal the most important and absolutely crucial features to discover knowledge and gain valuable information about hidden process characteristics and by this to foster future process and product developments in SM. Most effort is spent to constantly improve the prediction performance and optimize the ML models in terms of robustness, reliability of outlier detection and accuracy yielding results which meet highest demands with a relative deviation from the target between 0.5 % and 1 %. Even though many developments in VM achieved remarkable results, up to now no complete FS approach enabled the incorporation of all aspects and hence to solve all challenges and problems related to advanced VM in SM.

4.3 Feature Selection

After the fundamentals of FS as well as the challenges and the state of current research regarding FS application for VM are given in section 3.3 and section 4.1, respectively, the result of comprehensive research based on a literature survey is outlined in detail in this section focusing on most relevant FS with SVR, RFE and GA.

Feature Selection, Projection and Compression

At first, FS itself needs to be clarified and distinguished from other related but different dimensionality reduction techniques dealing with projection or compression. Projection approaches transform the variable input space into a feature space and introduce a ranking of the transformed features by their impact regarding predefined functions or rules. So-called PCA is designed to preserve intrinsic information while reducing present redundancy due to correlation of data [139], [149]. A major drawback is the preservation of all features in the projection which does not improve the computational performance [49]. Compression approaches reduces the variable input space by encoding information to achieve either lossy or lossless data compression using information theory [139], [140]. FS approaches are dedicated to select a feature subset of the entire variable input space not altering these features during the process of selection thus preserving the original structure and semantics of these features [81], [145]. Apparently, by maintaining the original features in the selected feature subset including their relationship to each other with hidden and unknown interrelations, domain experts are enabled and encouraged to interpret and discuss the outcomes (i. e. the selected features) regarding the reason why these features and their interrelations are selected to gain valuable knowledge about the investigated domain [139]. Hence, projection and compression methods are not suitable since almost no valuable process knowledge is revealed.

Heuristic Search for NP-hard Problems

As for many optimization problems and already stated in section 3.3, FS mostly belongs to the group of NP-hard problems as state space search to find optimal feature subsets inevitably resulting in a heuristic search [81], [12]. Regarding the decision how to start the heuristic search, forward selection with no initial features, backward elimination with all features and a randomly or seeded feature subset selection are possible options. Moreover, the decision of how to search in the search space needs to be specified by many available strategies whereof the most popular ones are already highlighted in section 3.3. A further decision is required on how to evaluate the feature subsets where predominant criteria are based on information theory or direct measurement of accuracy. At last, the stopping criterion for the search has to be defined commonly as a specified threshold for the resulting accuracy, a defined number of features left in the subset or simply a sufficient number of computed iteration steps [12].

Dimensions of Features and Instances

On the one hand, FS is often facing the problem of large variable input dimensions while on the other hand small sample sizes exacerbate the challenge to find the feature subset minimizing either the prediction error or the number of features (cf. section 3.3) [139]. Nevertheless, during the last decade the application of FS techniques and the investigation of the available variables changed from an optional preprocessing step to a real requirement prior to develop sophisticated DM or ML models [139].

Relevance of Features and Instances

Many irrelevant features and samples are present in most data and two challenges related to feature relevance arise as to determine which features to use and how to combine those features.

The amount of required instances for training combined with the number of features present in the DS to achieve a desired level of accuracy is described as the sample complexity [12]. In general, feature relevance is defined as a different outcome of the prediction if only the values of one feature differ and the remaining values of all other features are kept constant in at least two instances [12]. A drawback of this definition is the lack to recognize if a feature is truly relevant in case of too little instances are available for approval or rejection of the hypothesis. Strong and weak relevance of features are further defined and serve mainly for theoretical analyses of learning algorithms [12]. The presence of many irrelevant features combined with interactions among those features is shown to cause severe degradation of prediction performance (e. g. accuracy) for many FS algorithms (e. g. DTrees, Naive-Bayes classifiers) which are discriminating among classes independently of the amount of available instances in the DS [12], [78], [83]. In fact, Naive Bayes classifiers are more robust against irrelevant features but they are prone to correlated features due to their monotonic nature [81]. While most current ML approaches assume monotonicity of prediction performance, many real world scenarios do not satisfy these monotonicity assumptions due to the presence of irrelevant features resulting in poor prediction performance of for example DTrees [177].

Results of FS approaches are intended to be analyzed and used by domain experts as well as to be used for further data processing or modeling and differ from goals of feature weighting methods which are commonly performance driven and tend to be easier to implement. In feature weighting internal feature weights are optimized according to a specified function with least-mean squares and NNs as popular examples [11], [12] and extension for instance-based learning (e. g. nearest-neighbor) [163], information-theoretic metrics [30] and extensions to wrapper techniques to search through the weight space [82].

Similar to features, the relevance of each instance can be assessed by computationally intensive induction techniques learning faster with fewer instances, due to expensive instances labeling (e. g. expert analysis for every label) and improvement of the learning rate and the prediction performance by focusing on instances with a high level of information [12].

Feature Selection Methodology

Various conclusions in terms of finding the best feature subset are drawn in literature as referenced in the subsections above. Feature aggregation and construction prior to any FS is stated to possibly improve prediction performance in some domains. Nevertheless, the challenge which features to combine is similar to find an optimal feature subset. Two ways to tackle the problem of FS are suggested in [48]. On the one hand, investigation of a filter approach based on correlation coefficient or mutual information. On the other hand, a wrapper method with forward, backward or multiple feature selection should be considered to find optimal feature subsets [48].

Filter Methods

As an established and approved filter method, the RELIEF algorithm performs statistical analysis in linear time with the number of features and instances. Even though, no prior expert knowledge is required and no assumptions for the distribution of irrelevant feature values are required, the technique is stated to be unaffected by feature interactions and furthermore "fairly" noise-tolerant which comes in handy since feature relevance of the RELIEF technique is corroborated by theoretical analysis. In terms of feature reduction the RELIEF method does not necessarily find the smallest feature subset but noticeably shrinks the feature space where the final optimal feature subset can be obtained by exhaustive search conducted on the outcome of RELIEF. Another alternative to exhaustive search is the most common and subsequent application of induction learning methods with DTrees as popular example. Interval estimation can be used to discover the required threshold to discard irrelevant features statistically [12], [78]. Technically RELIEF finds the nearest hit and nearest miss close to each instance with commonly Euclidian distance as measurement for this nearest-neighbor approach and subsequently optimizes the feature weight vector. Generally, it can be stated that positive entries indicate more relevant features whereas features with negative weight vectors can be omitted [48], [78].

The FOCUS algorithm developed in [1] also minimizes combinations of features to optimize the discrimination among the classes. From initially isolated features, the filter technique starts to combine these features into pairs, triples and so forth until pure partitions are found for which no instances have different classes and then finally passing these feature subsets to a DTree as prediction method. The accuracy of the DTree degrades significantly by introduction of irrelevant features even though the FOCUS algorithm is not affected [12]. The FOCUS and RELIEF FS methods are also used with nearest-neighbor retrieval [15] and naive Bayesian classifiers [85] as induction methods. Similar variations of these methods are listed in [12].

Wrapper Methods

Noticeably, the application of SVM and SVR as pure wrapper learning methods for FS became very popular within recent years [148]. DTrees like ID3, C4.5 and CART are FS wrapper techniques with high risk to fail to exclude irrelevant features which would improve prediction performance whereas these FS wrapper methods yield computationally more efficient ML models compared to preprocessing FS filter methods [68], [81]. Other greedy backward elimination wrapper approaches are implemented with nearest-neighbor induction algorithms resulting in good classification performance while containing irrelevant features and fewer instances [12]. More research and application are performed on classification than numeric prediction i. e. regression (e. g. [164] with a k-Nearest Neighbor algorithm) [12]. Furthermore, naive Bayesian classifiers as approaches sensitive to redundant features and vulnerable to correlated features are shown to improve prediction performance when used in FS wrapper methods [86], [81]. Another investigation identified the equivalence of FS techniques regarding subsequently adding or removing features to or from the feature subset [132]. Also, an example is presented where a filter approach failed to achieve a correct prediction whereas a wrapper is able to deal with the dataset [81].

4.3.1 Feature Selection with SVM and SVR

Within the last two decades, SVM and SVR are investigated to optimize feature sets in the context of FS wrapper techniques partially with a kernel function to overcome the restriction to linearity. Dimensionality reduction of input space improves the prediction performance and speeds up SVMs inducing various research in this field with the state of the art approaches given below [173], [8].

As a first major contribution to FS with explicit SVM, the objective function as expectation EP of the error probability E of the ratio between all training data belonging to the squared set of size R and the corresponding squared margins M is minimized for the present optimization problem W by gradient descent [173]:

$$EP_{err} \leq \frac{1}{l} E \left\{ \frac{R^2}{M^2} \right\} = \frac{1}{l} E \left\{ R^2 W^2(a^0) \right\} \tag{4.1}$$

where α denotes the l Lagrangian multipliers. For large feature sets very small values of EP can be set to zero at once thus eliminating multiple features which can be repeated several times to speed up the process of otherwise greedy backward elimination. Superior performance compared to filter methods is shown in experiments with real-life data and toy data. This FS wrapper method with SVM as induction algorithm overcomes identified difficulties of SVM in high dimensional spaces with many irrelevant features [43], [173].

A second penalty-based method for FS linked to SVM is introduced with the focus on forcing a large number of weights to zero. A linear optimization is implemented and applied but unfortunately no concrete results are provided [49].

In comparison with filter methods, feature weights learned from an induction algorithm and feature ranking coefficients describe both the value of a feature and can be used to some degree interchangeably. Nevertheless, as a discriminant function using mutual information between features and based only on support vectors, SVMs provide a better feature ranking than correlation coefficients which favors SVR as technique to investigate regarding the present challenge [49], [81]. Also computational advantages of SVMs compared to other competitive learning methods are presented [29].

According to [108], embedded FS with SVR based on Kernel Penalization (i. e. KP-SVR) shows superior performance compared to other wrapper or filter based SVR due to better adjustment to data by optimization of the kernel function and simultaneous feature subset selection for regression. An equivalent computational effort compared to SVR-RFE is obtained and at least the same computing time as for backward elimination is required. A gradient descent approach is used to penalize features by finding the best suitable RBF-type kernel function within each dimension by combination of FS, generalization and goodness of fit. In addition to SVR-RFE, a search for the least relevant feature and a stopping criterion are implemented [108].

Inductive and transductive SVMs are designed and together with SVM trees, they are compared to each other on small, medium and large DSs where transductive SVMs perform at least as well as inductive SVMs on small and medium sized DS. SVM trees outperform the others on large DSs [138].

In bioinformatics SVR is combined with FS for multivariate calibration to identify amino acids (Phenylalanine/Phe, Tyrosine/Tyr & Tryptophan/Try) in their mixtures by fluorescence spectroscopy. According to [96], SVR achieves a better prediction performance in regression than NN and PLS for results of the LOO method. In addition to the assessed prediction performance in terms of accuracy (i. e. RMSE & MAE), FS for SVR is investigated by the filter method based on Mutual Information FS (MIFS) and the newly introduced embedded method Prediction RIsk based FEature selection for support Vector Regression (PRIFER). The fluorescence light intensities of 23 instances for 13 selected wave lengths as features are measured aiming to optimize the necessary features along with the accuracy for the three amino acids. The embedded FS technique PRIFER outperforms the filter technique MIFS by selecting only 9/4/9 features compared to 13/7/13 features for Tyr., Try. and Phe., respectively. The corresponding RMSE/MAE yields 0.17/0.11 for PRIFER and 0.19/0.13 for MIFS [96]. Even though the FS methods with SVR are conducted in bioinformatics with a different scope regarding the amount of considered instances and features, they motivate the investigation of SVR compared to NN and PLS. Neither any evaluation of outlier detection nor solutions for efficient and scalable deployment of this method are given in [96]. The 13 preselected features provide only little insight how well knowledge discovery can be accomplished. In the end, the improved prediction performance of a technique incorporating the prediction algorithm vs. a prior applied filter approach corroborates the significance of further research on wrappers and embedded methods to achieve highest prediction accuracy.

Summary SVM & SVR: Since due to differing experiments the results of the outlined publications for FS with SVM and SVR are individually hard to compare, a general assessment is provided. FS by means of SVM and SVR demonstrates superior performance for feature ranking compared to filter methods. Furthermore, first promising results are achieved for application to regression problems even though only little research is conducted so far using SVR. In general, the investigated techniques are due to their deterministic nature more prone to find local optima than heuristic ML methods based on the empirical risk minimization principle including random walk and are not focusing on outlier detection as well as efficient and scalable deployment in a productive environment. Also, the method is unable to deal with multiple selection criteria (e. g. concurrent optimization of the number of input features and prediction accuracy). Finally, no FS for subset optimization is considered but some approaches are motivated to reduce the feature space faster and more efficiently.

4.3.2 Recursive Feature Elimination

"A good feature ranking criterion is not necessarily a good feature subset ranking criterion." [49]. With this quote a first approach for SVM-RFE is introduced in 2002 as enhancement of backward elimination. The different challenges of finding adequate feature ranking criteria and feature subset ranking criteria is highlighted and an iterative procedure is given as solution formulated as RFE [49].

The strictly recursive and thus greedy approach is often not feasible due to an enormous computational effort which requires to turn the feature ranking into feature subset ranking by removing chunks of features at once. Interesting and important to emphasize is the fact that the top ranked features within a feature ranking are not necessarily the ones that are most important [49] which is also shown in later experiments in section 8.2. Hence, only the subset of the correctly selected top ranked features can yield an optimal feature subset.

The SVM-RFE algorithm reduces the number of initial features until the feature list is empty. For each passed step within the loop, the SVM classifier is trained and all α_i are calculated. Subsequently, for each feature within the high dimensional space the weight vector of the SVM is computed for all instances as sum over the products of the target value y_i and $\alpha_i \cdot x_i$ according to the support vector expansion in equation (3.11).

$$w = \sum_i \alpha_i y_i x_i \tag{4.2}$$

To square the computed weights is one way to obtain a feature ranking criterion which indicates the worst feature to delete from the feature list and to add to a ranked feature list. Hence, the computational effort increases linearly with the number of features. Superior performance of recurring RFE compared to one-time calculated FS ranking methods is also shown [49].

SVM-RFE is also compared with two other multivariate linear discriminant functions. At first, the Fisher linear discriminant function which is also called Linear Discriminant Analysis is based on orthogonality assumptions and solves a generalized eigenvalue problem. As second, the MSE linear discriminant function is computed by Pseudo-inverse. It is shown that SVM performs best or equal with already a small number of features [49]. Recent research on RFE confirms the superiority of RFE over zero norm FS or mutual information FS [47].

As fastest method following correlation techniques the computed weights w of SVM can be used as feature ranking criterion. For linear SVM-RFE, the coefficients α_i are optimized only once. To compute the weight vector β the support vector expansion is calculated as given in (3.11). Thus, for various feature subsets x_i only the sum of the product of x_i and $(\alpha_i - \alpha_i^*)$ needs to be computed yielding the computational advantage. For the nonlinear case using a kernel, only the kernel matrix H has to be recomputed for retraining whereat the quadratic time can be halved by copying the calculated values along the diagonal. Additional accelerations can be achieved by subtracting the partial scalar products of the eliminated features of matrix H and caching the already computed dot products. The complexity of SVM increases linearly with the number of features as seen above and quadratic with the number of available instances as the H matrix is of quadratic nature. For RFE equally useful and correlated features are not removed from the feature subset preventing loss of information and degradation of prediction performance. Although these features may not be ranked uniquely, one of these features is ranked higher and the other still remains in the feature subset whereas a naive ranking may produce equally important features which are then lower weighted themselves. It is already pointed out that RFE is more robust to overfitting (cf. section 3.3.4) than combinatorial search and other methods. In the end, the final number of selected features is not dictated by feature ranking methods [49]. SVM-RFE based on the approach presented in [49] is applied for diagnostic classification of mammograms where the FS algorithm is enhanced to a maximum margin minimum redundancy concept and compared to similar techniques in various ways taking into account redundancy and relevance. The induction method performs superior to the compared FS methods [77]. Higher prediction performance of SVM-RFE FS vs. correlation and information based FS approaches is also shown in [99]. In order to assess the stability of a feature subset in terms of the same selected features in several runs, SVM-RFE and Random Forest Variable Importance Measures provide different behavior. While the latter aggregates stable subsets, it includes not only the most important features whereas SVM-RFE finds the most important features and smaller feature subsets but suffers from a possible impact of the described imbalance rate (i. e. class samples versus intrusion samples). Improved understanding of process knowledge is achieved and stability of not changing feature subsets is identified to be an aspect to consider [95].

The crucial importance of DP for DM and Knowledge Discovery is corroborated in the case of SVMs where FS with SVMs critically depends on high quality data due to a strong influence of outliers [49].

RFE improves FS compared to other classifiers with best results by reducing chunks of features at once which loses impact for very high dimensionality vs. one feature at a time. Hence, without trading in accuracy for speed it is reasonable to eliminate chunks of features at the beginning of the FS process and refine to feature-wise reduction later on. As it is shown in experiments on the colon cancer DS SVM-RFE performs superior to SVM (cf. subsection 4.3.1) [49].

Finally, to optimize performance for FS a cutting plane algorithm is introduced with a layer to generate groups of features and then a second layer to select a group of these features with comparable performance for SVM-RFE [110].

Summary SVM-RFE: Various research with respect to SVM-RFE is conducted and clearly related to each other with predominant contribution according to [49]. Superior prediction performance is achieved by the recurring RFE approach and its additional elimination of chunks of features compared to one-time feature ranking techniques whereas still no real feature subset selection is performed. Neither application of RFE to regression problems nor focusing on outlier detection or efficient and scalable deployment in a productive environment are considered so far. As for FS with SVM & SVR no heuristic concept based on the empirical risk minimization principle is incorporated to explore feature space search via random walk and so prevent local optima. Also, the method is unable to deal with multiple selection criteria (e. g. concurrent optimization of number of input features and prediction accuracy).

4.3.3 Feature Selection with Genetic Algorithms

In order to overcome degraded prediction performance by violation of the monotonicity assumptions by inherent irrelevant features in case of DTree classifiers, GAs are introduced. In contrast to most proposed FS techniques, GAs are able to deal with multiple selection criteria as prediction accuracy, minimization of the empirical risk by avoiding local minima and the number of selected features in the feature subset. GAs are optimization methods and very effective FS wrapper algorithms in high-dimensional search spaces [177]. According to section 3.5 each individual in the population of every generation serves as a candidate for the feature subset of this generation and if inherited to the end for the final feature subset. Each feature is encoded as a single bit on the chromosome which represents a bit string of length equal to the number of features with an active feature indicated by a 1. Given m features a total amount of 2^m feature subsets exists and thus for a high number of features (i. e. m»30) exhaustive search is infeasible. The fitness function is encoded by the induction technique in the wrapper approach. Remarkable is the combination of several evaluation criteria in the fitness function namely the prediction accuracy and the computational cost [177].

A combination of evolutionary algorithms and SVMs is applied to optimize the number of features, the training error and the SVM model parameters. For the latter, the norm of the slack variables and the radius-margin quotient are tuned as dual objective which results in comparable models as using single-objective criteria [159]. A similar approach is already investigated two years earlier in the work of [148] with focus on optimization of classification accuracy and cardinality of the feature subset.

For using GAs as FS method to reduce the feature subset, it is shown that FS does not depend on the selected kernel function [160]. Further research with GAs and SVMs for time series classification is performed to underline no improvement of prediction performance but the achievement of the same prediction performance with a reduced feature set [34]. A comparison between FS by GA and SS is conducted in the area of computer aided diagnosis for computer tomographic colonography with SVM as induction method showing superior performance of GAs over SS [116]. The relative importance of each feature is assessed by using different runs of a GA technique [139]. Also, a combination of correlation-based FS with GAs is investigated. Initially, highly correlated features are grouped but subsequently only features with small correlation coefficients are used within the GA search to optimize the feature subset. It is concluded that this approach performs equally or better than various SS methods and some filter based methods [150].

GA as FS wrapper search algorithms is also used with SVM as induction method to predict clinical phenotypes based on genome-wide Single-Nucleotide Polymorphism profiles of sib pairs. 790 instances for 117 features are tested whereat the size of the feature set is halved in each generation until the classification accuracy improvement falls below a threshold of 0.001 and 18 features are left. The prediction performance of using GA FS with SVM as 'hybrid' is superior to a compared k-Nearest Neighbor 'hybrid' approach [44].

Substantial work considering several interesting aspects of related FS for SVMs and GAs is published 2003 [43]. At first, the generalization error is assessed by theoretical bounds instead of frequently used cross-validation decreasing the computational effort and improving the robustness against overfitting. The number of features and the SVM penalty parameter C are both encoded in the chromosome for concurrent optimization. Once more, FS for SVM is confirmed to significantly improve the prediction performance. It is distinguished between binary GA encoding where the number of features is known beforehand and decimal encoding with a varying number of features possible for the feature subset. As another result, the generalization performance of RFE and GAs using the R^2W^2 (cf. equation (4.1)) boundary implies an opportunity to save time if the number of features is known beforehand [43].

A similar GA approach for regression is chosen to optimize all considered SVR model parameters (i. e. $C, \epsilon, \gamma, \lambda$) where a mixture of a polynomial kernel and a RBF kernel with model parameters λ and γ, respectively, is applied and again concurrently the number of features are optimized. Thus, the chromosome bit string is encoded by the four SVR model parameters and all features. In the end, an improvement of the prediction performance can be shown [97]. Another work motivates the application of GAs instead of grid search to find the best features as well as the kernel parameter of a RBF kernel for SVMs. A ROC (Receiver Operating Characteristic) curve of sensitivity and $1 - specificity$ displays the benefit of a GA-based approach with fewer features [56]. Linking the GA search for features with model parameters achieves superior classification accuracy compared to single GA applications to either find the best feature subset or to only optimize the hyper parameters [170]. Furthermore, a weighted SVM is presented with GA-based parameter selection and improved accuracy [137]. Remarkable results for three extensions of GA-SVM, SVM-RFE and Recursive-SVM are observed in [98] for four DSs containing only few instances and features (both «100). On the one hand, Recursive-SVM and SVM-RFE are comparable in prediction performance (i. e. $0.017 < \text{MSE} < 0.186$) whereas GA-SVM performs noticeably better (i. e. $0.008 < \text{MSE} < 0.022$). On the other hand, the computation time of the GA-SVM technique considerably increases compared to Recursive-SVM. The conclusion states the fact that Recursive-SVM performs superior for classification but GA-SVM performs superior for regression [98].

Summary FS with GA-RFE: While the two previous methodologies are based on the principle of structural risk minimization of SVM, GA FS is based on the principle of empirical risk minimization thus encountering the challenge of overcoming local optima. FS by GAs demonstrates efficient wrapper algorithms in high dimensional search spaces and is independent of the kernel function of SVM. The application of GAs to regression problems is again not in focus of current research. The advantages of GA FS to perform feature subset selection comes along with the drawback of the lack of an individual features assessment which is highly desirable for smaller feature subsets obtained after significant reduction in the end of the process of FS. Finally, a high computational effort is required to optimize large feature sets by GA FS.

Summary: The advantages and requirements of space dimensionality reduction techniques to optimize the prediction performance and to minimize the feature set are clearly pointed out in state of the art literature. Superior feature ranking by SVMs is reported compared to correlation-coefficient-based methods or discriminant functions using mutual information theory. However, the highest scored features of a feature ranking approach does not necessarily yield the best feature subset and can omit complementary important but lower ranked features. Moreover, correlated but also useful features are often removed in correlation-based methods (e. g. filters). Improvements in terms of prediction performance compared to other FS techniques are shown by SVM-RFE with additional robustness against overfitting. Application of GAs improves determination of ML model parameters as well as feature subsets for SVM. Even though GAs can access multiple selection criteria (i. e. accuracy, number of features, minimization of empirical risk) no specific kernel selection is required. Thus, so far no suitable FS method is presented to unify the benefits of computational efficient feature ranking techniques for fast feature elimination and heuristic feature subset optimization constantly incorporating crucial interdependencies.

A highly interdisciplinary challenge is given to investigate and enable FS for regression in VM in the extremely specialized and complex SM industry. Within industrial manufacturing environment no dedicated experiments can be conducted and thus no assumed hypotheses can explicitly be proved or disproved. Generic VM is demanded for efficient and scalable corporate-wide VM implementation to encounter physically interrelated processes with hundreds up to more than 10.000 features with thousands down to less than 100 instances for a huge variety of logistical characteristics (e. g. recipes, operations, technologies, products, basic types, process groups). In conclusion, most research in FS deals with classification problems and so far less research is performed to investigate the challenge of regression for the apparently wide numerical range of features and instances to deal with regarding a corporate-wide implementation of VM. Finally, no approach reveals unexpected crucial features in terms of contribution to knowledge discovery in the area of VM in SM.

Before the newly invented smart FS algorithm to meet the outlined challenges and to solve the stated problems for advanced VM is described in detail in chapter 6, the requirements for an efficient and scalable implementation as well as the newly developed advanced VM system at Infineon itself are subsequently specified in chapter 5, hence, both together comprehensively explaining the scientific novelty of the present thesis which is finally capable to enable efficient, scalable, revealing and accurate VM in SM as outlined in the ensuing chapters.

5 Advanced Virtual Metrology System

The first part of this chapter considers relevant tasks to enable smart Feature Selection for an advanced Virtual Metrology system. Subsequently, the implementation of the newly developed advanced Virtual Metrology system in the productive environment of High Density Plasma Chemical Vapor Deposition at the Infineon frontend manufacturing site Regensburg is described in detail.

5.1 Enabling smart Feature Selection for advanced Virtual Metrology

In order to enable smart FS and by this an advanced VM system, a productive VM application including connections to all other related systems has to be implemented within an industrial environment to approve of entire concept regarding the challenges stated in subsection 4.1.1. Furthermore, the VM related systems R2R and FDC as well as the necessary data flow are outlined for better understanding of the productive advanced VM system.

5.1.1 Motivation and industrial Requirement for advanced VM

The immense competition within SM industry enforces economic efficiency (cf. challenge 1 in subsection 4.1.1) and targeted development of VM as APC application inevitably including conduction of a Cost-Benefit Analysis (CBA) as well as calculation of the expected return on invest. Challenges and problems dealing with the economic efficiency of corporate-wide VM deployment have been considered in various SM fabs and for several process areas. A state of the art literature review regarding economic benefits further motivating VM is provided in appendix A.2. The economic efficiency of VM implementation in the sense of maximizing the return on invest is corroborated by the result of a detailed CBA for the present use case in HDP CVD given in appendix A.3. The required framework to deploy the newly developed advanced VM system corporate-wide including the new smart FS algorithm is already available for R2R solutions. So, only the costs for human resources to develop VM have to be accounted. For the use case of HDP CVD four benefits are summing up to finally yield the total benefit of VM implementation taking into consideration 1) the reduced required metrology with 2) the derived improved cycle time, 3) less possible scrap production of a subsequent planarization process and 4) the higher production output due to improved utilization of the already installed metrology tools avoiding operational bottleneck scenarios.

5.1.2 Virtual Metrology related Systems and Data Flow

Various established systems and applications are related to VM and necessary prerequisites for its implementation at Infineon. The FDC and R2R control systems as well as an Infineon internal framework to develop and deploy MATLAB-based applications in combination with SM fab interfaces and an enterprise DB are highlighted in the following.

Fault Detection and Classification System

A FDC system is designed to access and monitor equipment parameters stored in an online database system during production to detect and classify out-of-control equipment states responsible for possible misprocessing of wafers and thus avoidable scrap. Various online reactions such as notification to process engineers or instant tool stop can be triggered to ensure highest possible control. In detail, tool-integrated as well as additional sensors provide time series data for hundreds of physical process parameters like the already mentioned process chamber pressure, applied radio frequency power or induced voltage. In order to reduce the stored data volume and thus, also enabling historical long-term analysis of process conditions, the time series data are aggregated by calculation of relevant values (e. g. mean, standard deviation) for the individual process steps. Furthermore, higher aggregation is performed by algebraic combination of the aggregated process parameters to additionally generate significant process relevant parameters for FDC application. Based on these process parameters, limits for various combinations of logistical information (e. g. product, recipe) can be set and the type of reaction in case of process parameters exceeding these limits can be defined. Upper Control Limit (UCL) and Lower Control Limit (LCL) are sufficient in the scope of this work and define a control range for the individual process parameters to ensure stable process conditions. Finally, these aggregated process parameters serve as input variables, so-called features, for the ML algorithm to predict the process outcome as output of the VM system.

Run-to-Run Control System

A R2R control system is established to enable direct process control by adjustment of relevant recipe setpoints on the individual production equipment resulting in appropriate adaption of essential process parameter settings to ensure the desired process outcome. Prior the processing of productive wafers, the specific recipe setpoints actually valid for the logistical characteristics of the associated lot (e.g. product, basic type) are downloaded from the R2R controller to the production tool. After the lot is processed with these predetermined parameter settings and proceeded to the subsequent metrology operation, the physical measurement result for the sampled wafers is uploaded to the R2R controller and compared to the internally calculated target value from which the initial lot-specific recipe setpoints were derived. Depending on the resulting deviation between the calculated and thus expected process outcome and the real measured, the internal regression model parameters of the R2R controller are updated accordingly. Typically, Exponentially Weighted Moving Average based models are used for R2R control. As a result, process drifts or offsets caused by adjustments of preceding processes, degradation of the production equipment or relevant maintenance interventions (e. g. process chamber wet cleans) can be compensated by deploying such closed-loop R2R control. As a consequence, the process parameters adjusted by the R2R controller need to be considered accordingly for VM modeling in order to cope with related parameter drifts and offsets which does not affect the process outcome.

Infineon Framework to deploy Virtual Metrology

Many commonly used SM fab IT systems including their interfaces are exemplarily outlined in [18]. In order to enable fast and agile development and implementation of VM with respect to selected use cases, an Infineon internal framework is used allowing 'plug and play' of any type of MATLAB-based algorithms. This framework, which is also dedicated to integrate the R2R control system into the Infineon fab environment, is connected to various other IT systems and also provides interfaces to necessary enterprise DB. All connections and the data transfer can be performed by specific method calls.

Data Flow

Nowadays high-end SM requires enormous data processing for each equipment to set recipe setpoints and their equipment-wise individual adjustments with the challenge of older tools containing older microprocessors resulting in lower bit rates available for data transmission. Due to the fact that crucial instructions are executed with highest priority, a lack of additional read-out data (e. g. for analysis or development as in the present case) during online manufacturing is possible and can complicate VM development. Received process data are further preprocessed and together with logistical/context data stored into the R2R- and FDC DB. Even more data are obtained by other data sources and also stored to various DBs or data warehouses. Many departments (e. g. industrial engineering, product technology development & chip design) are focusing on different tasks using all available data sources. Thus, the complexity of the entire data flow with all interfaces, DBs and enterprise servers cannot be covered in detail within the present thesis.

However, all data expected to contain useful information for VM are collected from a central data warehouse, an engineering DB and the R2R- and FDC DB. Process data measured by integrated equipment sensors are queried and received from the tool as continuous data stream. The extensive time series raw data stream is stored into fragmented files on a file server which are then further processed. Finally, any predefined and calculated key numbers (e. g. mean, variance, functions of other key numbers) are stored within an enterprise DB. The challenge to handle the resulting data fragmentation combined with missing data underlines the complexity of implementing a VM system.

5.2 Advanced Virtual Metrology System at Infineon

Starting from scratch at the beginning of this work in February 2011, an advanced VM system has been initially investigated subsequently developed, implemented and finally tested and evaluated online within the previously described Infineon internal framework. Initial activities focused on the investigation of applying a wide range of ML regression techniques to accurately and reliably predict the process outcome for several use cases for different processes (cf. subsection 4.2.2). As a promising use case with significant economic benefit (cf. subsection 5.1.1), the HDP CVD inter-metal dielectric process (cf. section 2.2) was finally chosen to develop and implement an advanced VM system. Compared to state of the art VM systems, the newly invented VM system is advanced by smart FS to automatically reveal crucial features and by this to achieve the high aims of enabling and optimizing efficiency, scalability, knowledge discovery and accuracy as the essential demands made on a corporate-wide VM system.

Figure 5.1 visualizes VM as an enhancement of physical metrology. Some of the production wafers processed on the manufacturing equipment are sampled for physical metrology and the measurement results are monitored via statistical process control with statistically calculated upper and lower control limits (UCL, LCL) as displayed in the upper part. As not all productive wafers are inspected, possible misprocessing of unmeasured wafers remains undetected and a breakdown of the finally manufactured devices reducing overall yield implying less efficiency and thus higher manufacturing costs per device. The lower part shows the integration of VM into SM where process sensor data and metrology data serve as input for supervised learning DM models (e. g. NN, M5', SVR). As sensor data are available for all processed wafers, the process target can be predicted for every wafer and monitored in the same statistical process control chart. Several advantages as less real metrology, savings of materials, production time and equipment degradation as well as increase of product quality are well-known (cf. section 2.2, appendix A.2). An ancillary but important effect is the availability of the VM prediction as input for the R2R controller enabling Wafer-to-Wafer instead of Lot-to-Lot control as it is realized nowadays.

5.2.1 Knowledge Discovery and Data Mining

State of the art research in the field of VM mainly focused on evaluation and tuning of regression and induction algorithms as well as basic FS methodologies (cf. section 4.2). In order to explore and unleash the full potential of structured Knowledge Discovery in Databases and DM (cf. section 3.1), the entire CRISP-DM approach has been investigated, adapted and enhanced for VM at Infineon.

Business Understanding

The economical motivation to implement a VM system is already highlighted in detail above (cf. subsection 5.1.1 & appendix A.2). Nevertheless, the effect of Knowledge Discovery in Databases to improve the knowledge about highly complex processes and thus to serve as key

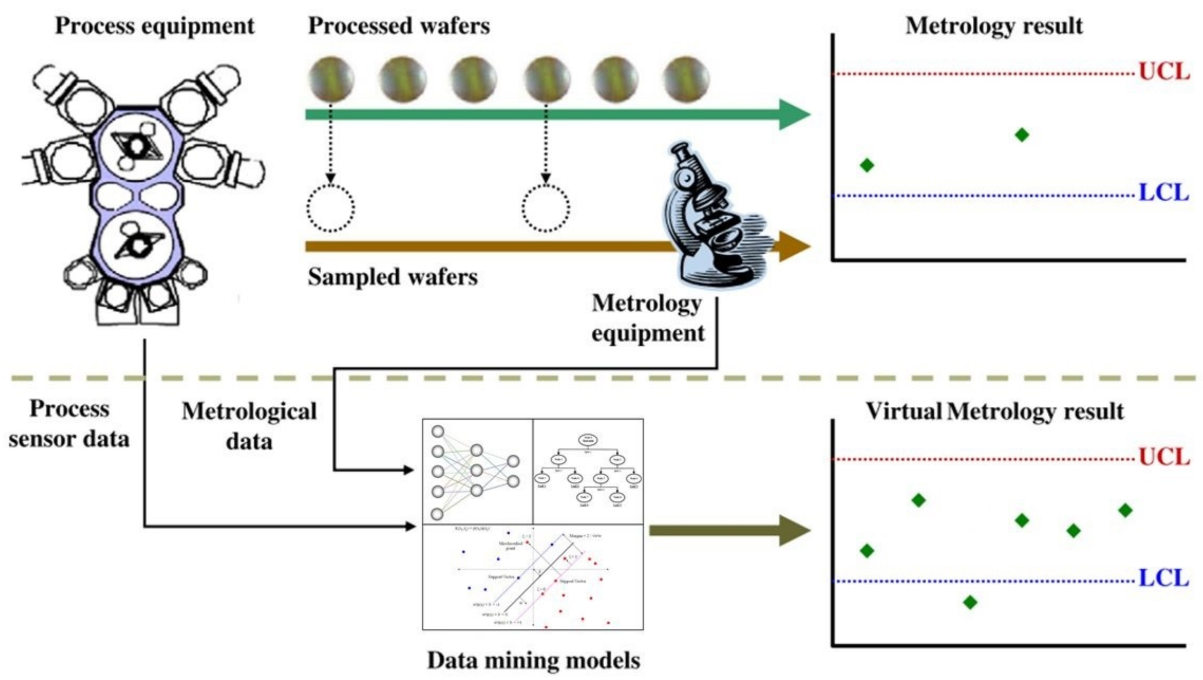


Figure 5.1: Overview of Virtual Metrology as enhancement of physical metrology (cf. [70]).

enabler for future unit process developments, supporting future technologies and products is not comprehensibly considered so far. Hence, process experts of the Unit Process Development CVD team were involved as stakeholders. The substantial challenge of enhancing their profound knowledge about the HDP CVD process by means of Knowledge Discovery in Databases is encountered by smart FS. So, the process of business understanding identified the challenges of economic efficiency, scalability and accuracy as basic requirements and knowledge discovery as ambitious goal to achieve by advanced VM.

Data Understanding

Within a productive manufacturing environment lots of data are collected in different DBs for various applications. First data collection was concentrated on data stored within the FDC DB (cf. section 5.1.2). In addition, a data warehouse was inspected to get further useful information which was successfully achieved by obtaining an additional parameter for approval of logistical information on wafer identification. This parameter was merged to the FDC and R2R data. Furthermore, another engineering DB was queried to obtain more logistical parameters enabling more specific tests due to an increased granularity of available data.

Critical tasks of the processing procedures are prioritized by the control software of the production equipment compared to services for data acquisition via the tool interfaces resulting in possible data gaps during data collection due to the limited processor speed of the equipment internal control module. Such data gaps within the time series data collection can yield missing values if a FDC process parameter cannot be calculated. Thus, dealing with missing data is inevitable even though the mentioned parameter from the data warehouse was useful to improve the data quality.

Correlation analysis was conducted to recognize characteristical interdependencies between process parameters but no data were removed to keep all possible information for later FS.

Correlated features could also be modeled by cluster analysis and instances with well-known characteristics (e. g. after maintenance action) could be illustrated, but no further effort was spent on the tasks due to the present supervised learning approach for VM versus unsupervised learning clustering techniques.

All variables of any DS were explored and inspected carefully but no initial hypothesis to verify or reject was put forward.

Data Preparation

- **Data Formatting**: Each logistical parameter which can be considered to be important to add valuable information is converted from character, string or date into natural numbers because numeric input is required for SVR.

- **Data Set Compilation**: All DSs are compiled using two years productive HDP CVD SM fab data.

- **Feature Translation**: The investigation over the past years consistently confirmed that the prediction performance is improved by using the Deposition Rate (DR) instead of the Layer Thickness (LT) as prediction target. The calculation of the DR as well as normalization of all data are the conducted tasks in feature translation.

  A R2R controller running on the respective production equipment calculates the Deposition Time (DT) based on an Exponentially Weighted Moving Average filter approach from the estimated DR and adjusts the DT recipe setting for each lot. Thus, the DT has to be excluded from the input parameters to avoid indirect modeling of the computed function of the R2R controller instead of approximating a function which precisely models the HDP CVD process. In order to exclude the DT from the input without any loss of information, the LT is divided by the DT. The obtained DR given in equation (5.1) serves as prediction target for the VM system but it excludes the DT as a major source of information for the prediction of the LT. Various ancillary experiments have proven a perfect correlation between DT and LT allowing to set this boundary condition for the present thesis. Further investigations regarding this issue are already discussed in [92]. The initial target Y (i. e. the deposited LT in nanometer [nm]) is measured at nine different sites on the wafer surface whereupon the mean is calculated to receive a robust target value Y independent of the uniformity of the deposited layer (cf. subsection 2.3.3). To test the precision of SVR, the DR was predicted and then translated into the LT by multiplying with the associated DT. Hence, for the finally conducted assessment of the resulting RMSE the LT is used.

$$\mathrm{D}R\left[\frac{\mathrm{n}m}{\mathrm{s}}\right] := \frac{\mathrm{L}T\ [nm]}{\mathrm{D}T\ \ [s]} \tag{5.1}$$

As second part of feature translation, an effective technique to avoid domination of large scaled input values over small ones is to normalize all variables to a specific range which is set to [0,1].

- **Instance Selection**: In order to ensure highest accuracy for VM, restrictive instance selection is chosen during the development where for any missing value (i. e. NaN) the entire instance is removed from the DS. No replacement by other calculated values (e. g. mean) is performed. All obvious outliers of any feature or target value turned out to significantly degrade the prediction performance of SVR [26]. Thus for the evaluation of the developed new ERBE FS algorithm (cf. section 6.5) all instances showing significant outliers exceeding the $3\,\sigma$ standard deviation range are removed.
- **Feature Selection**: In the scope of the present thesis, only features from the productive deposition step are considered. Features from the same process but not this specific deposition step (e. g. previously performed cleans, heating process steps) are not considered whereupon these additional features can contain valuable information for further inference [123]. All features containing only NaN values or only a single value are immediately removed whereas missing values are replaced by the mean of this feature vector.

Modeling

Within the last decade SVR as a powerful and promising induction method has evolved with the advantages of high accuracy and strong predictive power combined with good generalization ability. The multivariate regression technique was enabled by means of kernel extensions to deal with nonlinear data without computational drawbacks. If model robustness can be handled by smart data preparation and skilled model parameter adjustments SVR can also serve as reliable prediction method. Based on the principle of structural risk minimization and only linear increase of complexity with the number of features with further optimization potential of multiple feature-independent kernel extensions and adaption of Sequential Minimal Optimization (cf. [124]), SVR appears to be a promising method for highly accurate and reliable VM.

Various regression algorithms (e. g. PLS regression, CART, NN regression) were evaluated for VM with generally strong prediction performance whereas other techniques (e. g. MLR, Simple Linear Regression) showed limited or inferior performance (cf. section 4.2). A combination of the former methods (i. e. M5', NN) is evaluated and implemented in [87] and is used within the developed VM system introduced in the following subsection 5.2.2.

Evaluation

The importance to achieve highest required accuracy and to detect critical outliers yields the decision to choose the RMSE as main evaluation criteria due to the fact that higher deviations from the real metrology have higher impact to the final error. The common MAE is provided for reference. For both the coefficient of variation (i. e. CV(RMSE)) is calculated to obtain scale-independent results and comparability with referenced state of the art VM (cf. chapter 4). In addition to the single value error measurement, the quantitative assessment of revealed outliers is conducted by the sensitivity giving the ratio of detected outliers out of all outliers. The coefficient of determination R^2 as third complemental testing method defines how well the prediction model fits to observed data. All experiments are carried out to test the designed VM system and the newly implemented ERBE algorithm which reveals the most important process parameters.

Deployment

Within the scope of the present thesis, the developed VM system including the following new ERBE FS algorithm (cf. section 6.5) is implemented as online application and tested on real-world productive process data to corroborate the concept of the provided VM system enhanced by the good performance of the ERBE algorithm. Future deployment is planned but related activities are out of scope of this work.

5.2.2 Advanced Virtual Metrology System Implementation

An advanced VM system was developed and implemented within the productive SM environment at Infineon which is associated with the benefit of ensured availability of sufficient real production data but also with the challenge to collect experimental data necessarily without any disturbance of the running manufacturing. Compared to a pure academic approach, a comprehensive amount of enlightening experimental data (e. g. outliers for specific settings) could hardly be incorporated into the investigation due to the missing opportunity to design experiments for specific targets to investigate because the entire manufacturing must not be affected or changed at any time. In addition to the advantages of real industrial data collection and disadvantages of a lack of comprehensive experimental data, other productive systems are influencing and altering the research and development activities (cf. subsection 5.2.1 and equation (5.1)). The implementation of the advanced VM system was accelerated by the concurrent Infineon internal development of a new agile master framework based on modular design principles to allowing rapid deployment of applications implemented in MATLAB. Interfaces to other fab systems (e. g. Manufacturing Execution System) and various DBs were provided together with load balancing and many other framework functionalities. Hence, efficient development and implementation of the present advanced VM system as an internal subsystem was substantially supported by the master framework. The VM system comprises independently running VM modules performing the prediction and training of implemented ML algorithms as well as the configuration in terms of logistical granularities (e. g. equipment, technology, product, recipe, operation) and available VM models.

Prediction and Training Module

The in the scope of this work developed Prediction and Training Module (PTM) module located as application within the previously mentioned internally and concurrently developed framework (cf. section 5.1.2) specifies an independently running instance of the VM implementation which

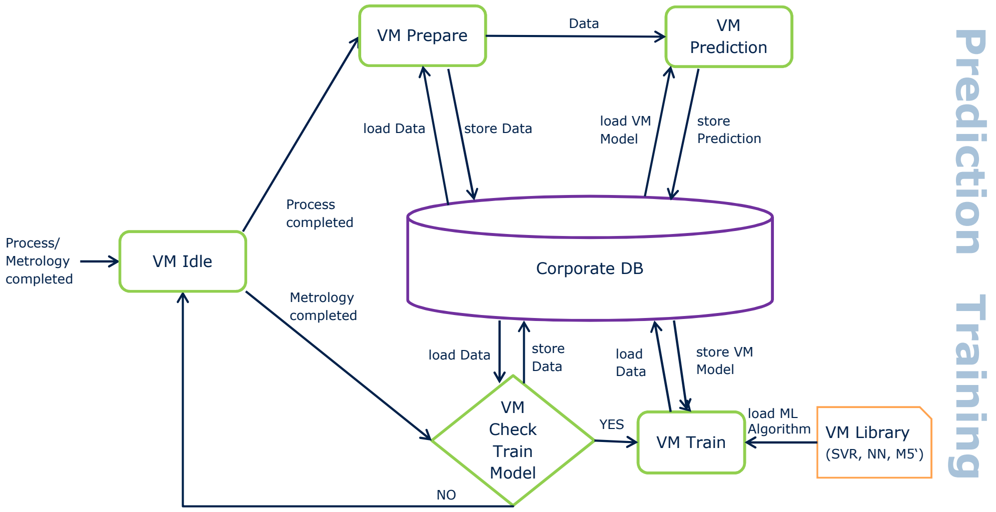


Figure 5.2: Advanced VM System Prediction and Training Module: The scenario of a completed process with a final prediction is displayed at the top whereas a completed metrology with eventual new training of VM models in case of inaccurate predictions is illustrated at the bottom.

can adopt different states depending on the provided input and the previously computed VM predictions. Figure 5.2 illustrates the PTM workflow diagram with five possible states (lime), the VM library containing SVR, NN and M5' as ML methods (burnt orange) and the aggregation of various corporate DBs (purple) including the relevant transitions and interactions (dark blue arrows). Two VM scenarios are processed to either calculate a new prediction or to train the VM models. For clarity the paths to return to the idle state after storing a prediction or the VM models are not indicated.

Prior to an activation of the VM system, the corporate DB has to be initialized with a basic dataset containing physical process, logistical and metrology data for a wide range of logistical granularities (i. e. various equipment, products, recipes, technologies). Already computed predictions are not mandatory to be included during initialization but predictions for the loaded instances performed by already trained VM models accelerate the processing performance of the VM system during the first weeks of operation due to a decreased effort necessary to train capable VM models if no predictions are available for already existing instances.

For both scenarios the VM module registers itself as listener in the framework to get notifications from the fab system. Thus, the VM implementation is encapsulated and other different fab interfaces are connected only to the framework which provides MATLAB methods to be called directly inside the VM source code. For the application in the productive environment, many independent instances of the VM implementation can be populated and configured in parallel to listen and to serve different processes and equipment, all accessing the same provided interfaces, DBs and libraries. The high potential of this new generic advanced VM system is unveiled by the fact that future activities can focus on extensions and enhancements of the advanced VM system and the adjustments of individual VM implementations while using the present implementation as a template. Future VM is enabled to run independently from human interactions by finding the most important features through application of the ERBE FS algorithm to achieve highly accurate and reliable predictions and by automatically triggering necessary VM model updates online during production.

In the following, the PTM workflow for productive application is described in detail.

Prediction: After completion of a production process for HDP CVD a notification is sent to the fab system. The PTM which is registered to listen to the associated process equipment will be triggered and basic process details in terms of specific logistical parameters for the processed lot or rather wafer (e. g. lot/wafer ID) are provided. These logistical parameters are kept and the state of the VM module changes to 'VM Prepare'. Depending on the revealed features as outcome of the FS, only the selected process features and the absolutely necessary fraction of all available logistical data to individually describe the present process are queried from the various DB sources to ensure most efficient VM processing. Subsequently, data are merged as well as preprocessed (cf. section 2.1, subsection 5.2.1) and the purified wafer instance is stored into a dedicated DB as well as passed to the 'VM Prediction' state. The appropriate trained VM models (i. e. SVR, NN & M5') are loaded from the configuration DB according to the specific logistical information as used before to query FDC data. Afterwards, the process result is calculated by the VM models which can be performed instantly for a single wafer instance. Finally, the predicted VM target is stored to the already present, purified dataset and the PTM returns to the 'VM Idle' state. Now, the predicted VM result is also sent to other systems for further analysis and statistical process control. The upcoming paragraph describes the estimation of the prediction reliability together with possible reactions.

A RI is suggested based on comparison of predictions performed by MLR and other learning algorithms with NN as popular example [24], [71]. Due to the fact that the inferior performance and reliability of MLR is already shown [127], a different approach is developed to access the reliability of a VM prediction in the scope of the present dissertation. Various high-sophisticated ML techniques (i. e. SVR, NN, M5' (cf. subsection 3.2.2)) based on different statistical fundamentals have been tested for accuracy and reliability with comparable prediction performance (cf. [93]). Hence, a traffic light logic is derived where at first the prediction is expected to be acceptable (green) if all methods compute a similar prediction, secondly the prediction should be treated with care (yellow) if the variance of the three predictions exceeds a defined range and thirdly the prediction has to be rejected (red) in case of deviating predictions above a certain limit of at least one method. Different prediction outcomes and behavior from these methods may result from their individual mathematical characteristics for underlying data (e g. distributions, shifts, drifts, kernels). For the red light case, a notification is sent to the process experts and a physical metrology is triggered as well as an inspection of the VM models in terms of recent prediction accuracy. Repeated deviating predictions cause a retraining of the VM models. The applicability of the suggested traffic light logic RI approach was corroborated by results achieved on productive data with all three methods (SVR, NN & M5') as illustrated in figure 5.3 from presented publications [87], [93]. In order to perform an in-depth assessment of the traffic light logic RI an evaluation over a longer time period is necessary and thus content of future work.

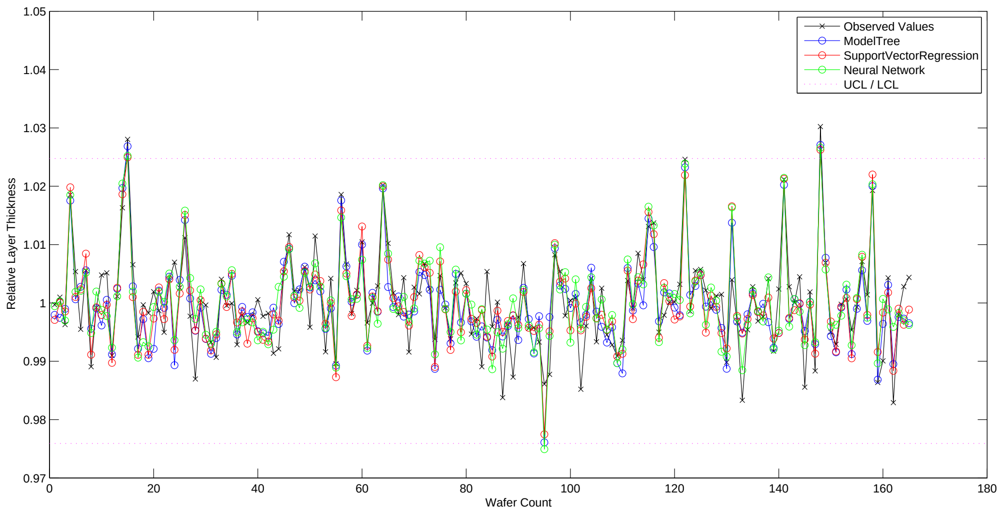


Figure 5.3: ML algorithm comparison for a VM Reliance Index: "Prediction by ML techniques: M5' (blue), NN (green) and SVR (red). The prediction performance is remarkable achieving a CV(RMSE) < 0.5 % and $R^2 > 70\,\%$. The deposited layer thickness as prediction target was measured outside the control limit for two wafers whereupon all algorithms recognized these wafers as outliers. Also, all other observed values close to UCL and LCL are fitted perfectly" [87], [93].

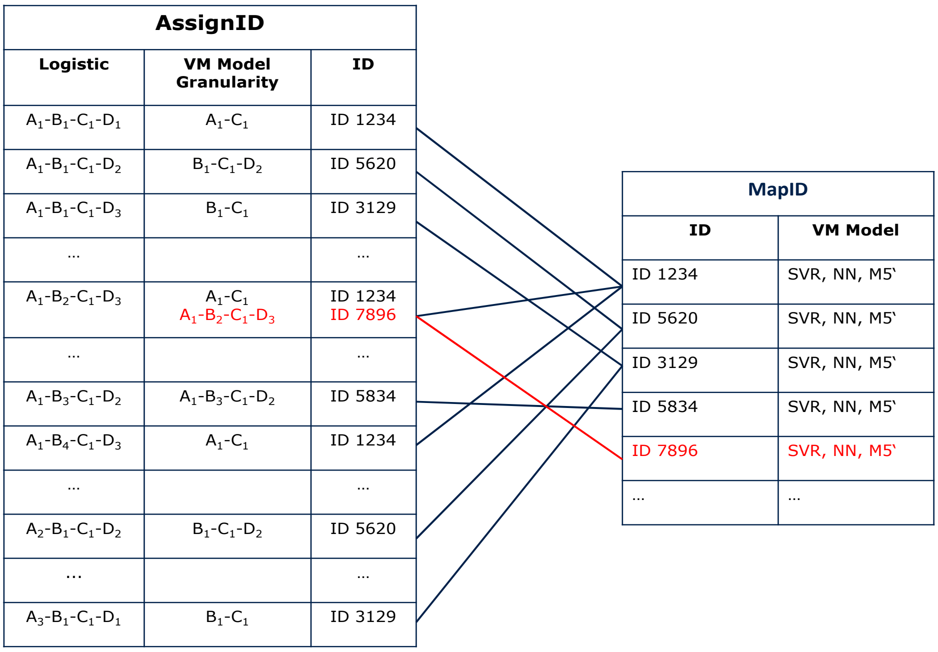
Training: After a wafer is measured, the listening VM module is triggered and required logistical parameters are passed to the 'VM Check Train Model' state. Subsequently, the corresponding measurement data and the already available purified dataset including the calculated prediction are loaded from the DB and merged. The VM prediction is compared to the real metrology result and according to a predefined logic, a decision is made to retrain the VM models or just store the dataset in the DB and return to the 'VM Idle' state. The predefined logic for a configured logistical granularity can be manifold and, for instance, may depend on the frequency of the used VM models whose reliability is to be accessed according to the prediction performance in the past. The challenge in low-volume-high-mixture SM is to find rules which imply enough generalization not to retrain the VM models to often which would cause instabilities or even oscillating predictions in case of short MW DSs as well as computational overload. But at the same time the rules need to be specific enough to ensure high accuracy by adequate frequency of the model training. A simple rule during evaluation of the implementation enforces a new training if a predefined limit for the deviation of the prediction from the target is exceeded by more than 20 % of the last (e. g. 10) predictions. An investigation of various scenarios as best time point to trigger a new training to achieve best accuracy and reliability is content of actual research (cf. section 10.2).

If a new training is necessary, only some information containing the logistical granularity need to be passed to the next state 'VM Train' because the metrology result for this instance is already stored in the DB. As first action in this state, the entire dataset for this logistical granularity is queried from the DB. After the purified DS for the training (e. g. MW approach) has been loaded, the available implementations of the ML techniques are loaded from the 'VM Library'. Finally, all three VM models for SVR, NN and M5' are retrained to ensure comparable predictions based on the same training DSs for different logistical sub-granularities which is described in further detail in the next paragraph below. The VM models for the individual logistical granularity yielding the most accurate result are selected as productive VM models for further productive predictions and stored into the DB before the PTM returns to the 'VM Idle' state.

Configuration Module

In low-volume-high-mixture SM a huge variety of products based on a lot of different technologies is manufactured on an individual production equipment. Thus, for just one specific process (e. g. HDP CVD) many different products and technologies including various operations and recipe settings are processed on each equipment during daily production. Even a more detailed categorization of this logistical granularity (e. g. basic type) is considered for VM (cf. [128]) but for simplicity reasons, to facilitate system efficiency by minimization of configuration and maintenance effort as well as for consecutive developments a solution minimizing the logistical granularity is highly desirable. Highly accurate and reliable predictions combined with computational efficiency can be achieved by individual adaption and optimization of the VM model according to the required logistical granularity. For more frequent manufactured products VM models can be tuned for a higher logistical granularity where data is separated by more logistical parameters since still enough data are available to reveal crucial characteristics whereas low volume products need VM models tuned for a lower logistical granularity to enable meaningful statistical investigations based on a sufficient amount of data. Accordingly, a Configuration Module (CM) has been developed for the advanced VM system to cope with this demand and to allow optimization of the prediction performance and reliability as it is required for a productive VM system.

VM models for a specific logistical granularity are trained on all partitions which can be created for the respective logistical parameters. Thus, the total number of partitions is given by the Bell number B_n where n describes the number of elements in the partition which are the different logistical parameters in the subsequent example. In order to prevent multiplication of trained VM models for lower logistical granularities, two tables are designed to efficiently assign the logistical granularities to the VM models.



AssignID

Logistic	VM Model Granularity	ID
A_1-B_1-C_1-D_1	A_1-C_1	ID 1234
A_1-B_1-C_1-D_2	B_1-C_1-D_2	ID 5620
A_1-B_1-C_1-D_3	B_1-C_1	ID 3129
...		...
A_1-B_2-C_1-D_3	A_1-C_1 A_1-B_2-C_1-D_3	ID 1234 ID 7896
...		...
A_1-B_3-C_1-D_2	A_1-B_3-C_1-D_2	ID 5834
A_1-B_4-C_1-D_3	A_1-C_1	ID 1234
...		...
A_2-B_1-C_1-D_2	B_1-C_1-D_2	ID 5620
...		...
A_3-B_1-C_1-D_1	B_1-C_1	ID 3129

MapID

ID	VM Model
ID 1234	SVR, NN, M5`
ID 5620	SVR, NN, M5`
ID 3129	SVR, NN, M5`
ID 5834	SVR, NN, M5`
ID 7896	SVR, NN, M5`
...	...

Figure 5.4: Advanced Virtual Metrology System Configuration Module: All possible logistical combinations are assigned to an ID (table AssignID) and mapped to the corresponding VM model (table MapID). In table AssignID, several different logistical combinations (1^{st} col.) can apply to the same VM model granularity (2^{nd} col) and uniquely assigned to a generated ID (3^{rd} col.). In table MapID these IDs are mapped to the trained VM models which are stored in the DB and thus not multiplied for several logistical combinations. In red, an example for a required retrained VM model is highlighted with changed entries.

Figure 5.4 visualizes a possible configuration of the VM system without the claim of being exhaustive. Four different logistical parameters (e. g. equipment, product, technology, operation) are represented by $A_i - D_i$. The index i indicates the different possible values of the logistical parameters (e. g. A_1 could represent product 1). The left table AssignID assigns the complete set of permutations for all logistics (1^{st} column) with corresponding VM model granularity (2^{nd} column) to a randomly generated ID (3^{rd} column). These IDs are then mapped in the right table MapID to the actually trained VM models (i. e. SVR, NN, M5’) which are stored in the DB. The VM model granularity is defined as the combination of the logistical parameters (i. e. $A_i - D_i$) to achieve highest accuracy with the training DS. So, more generalized VM models can be built on a combination of fewer logistical granularities (e. g. only B_1 & C_1 – ID 3129) including all data of not specified logistics (i. e. A_* & D_*) due to the fact that for lots with rarely manufactured combinations of logistical granularities (e.g. rare operation, technologies, products) not enough training and validation data are available if higher granularities are specified. The assignment of logistics to IDs in table AssignID evolves empirical over time and with necessary retraining of a new VM model a different granularity may be assigned for the new VM model to guarantee highest possible accuracy and reliability.

The VM models of the first example with identifier ID 1234 are only trained on data of the logistical granularities A_1 and C_1 whereupon the logistical granularities B_1 and D_1 are not considered (i. e. B_* & D_*). Hence, the same VM models are used for the logistical granularities A_1-B_1-C_1-D_1, A_1-B_2-C_1-D_3 and A_1-B_4-C_1-D_3. Analogously, more generalized VM models are trained for ID 5620 and ID 3129. An example for a highly specified trained VM model is provided with logistical granularity A_1-B_3-C_1-D_2 and ID 5834.

In case of degraded prediction performance for VM model of logistical granularity A_1-B_2-C_1-D_3 with ID 1234 (illustrated in red), new VM models are trained for several possible logistical granularities (e. g. A_1-B_2-C_1-D_3, A_*-B_2-C_1-D_*, A_1-B_*-C_*-D_*, etc.). Nowadays due to an increased demand of products and the resulting adaption in manufacturing more data are available for the specific combination of logistical granularity A_1-B_2-C_1-D_3. Thus, a more specific VM model can be built for exactly this granularity combination and yields the best prediction performance. A new ID (7896) is generated and assigned in the table AssignID. Subsequently the new ID 7896 and the trained VM model are added to the table MapID and stored into the DB and the old link will be deleted (not visible in figure 5.4).

The reduction of DB storage for different logistical granularity becomes obvious due to the fact that multiplication of the same VM models used for different logistical granularities is avoided (e. g. ID 1234). The size of the mapping tables is negligible compared to otherwise multiply stored VM models.

For productive online processing the CM is queried by the 'VM Prediction' state of the PTM to provide the dedicated VM models for the given logistical granularity. The CM is located within the corporate DB and therefore not specifically indicated in figure 5.2. An incoming request is routed via both CM tables and the trained VM models are returned to the PTM state to calculate the VM outcome.

5.2.3 Boundary Conditions

The accuracy of the layer thickness measurement by the metrology equipment as described in subsection 2.3.3 can be estimated according to the specification sheet [161] of the metrology equipment manufacturer Therma Wave to be 0.4 % RMSE deviation of the deposited layer thickness while precision and repeatability are stated as 0.05 % and 0.1 %, respectively. Hence, a VM prediction accuracy of smaller than 0.4 % cannot be achieved or evaluated since the prediction cannot outperform the metrology equipment on which underlying training data the VM model is built.

In addition, the target value for the liner sub-layer deposition (cf. subsection 2.3.3) is subtracted from the total metrology result in productive mode because the liner sub-layer deposition is a well-controlled, reliable and accurate process. Thus, the impact of the limited accuracy of the metrology tool on the final VM prediction is extended by a tiny variation and uncertainty introduced by the liner sub-layer thickness.

Summary: Following the state of the art of the previous chapter the new VM system is described in detail. The CBA as requirement to approve economic efficiency and VM related systems and data flow complete the comprehension for the introduced VM system. Initially, the important incorporation of the CRISP-DM process for knowledge discovery and DM is highlighted. Subsequently, the essential implementation of the advanced VM system is precisely outlined with PTM and CM as core pieces. Finally, the boundary conditions round off this chapter. The newly invented smart FS of the next chapter enables the implemented VM system to incorporate only the most important features and thus to significantly enhance efficiency, scalability, process knowledge and prediction performance. The VM system is improved to an advanced VM system as first totally generic and efficient VM approach and as first core piece of the present thesis.

6 Smart Feature Selection

Following the description of the VM system the advancement by smart FS is subsequently outlined in detail. The approach of the present thesis to optimize the feature subset is based on and corroborated by upcoming experimental results demonstrating the benefit of ERBE for VM in SM. As it is shown in [49], the selection of the most important features has a significantly higher impact on the prediction performance than the choice of comparable learning algorithms which is affecting the result only marginally. Furthermore, FS was recognized as a mandatory approach to achieve good results in case of high input dimensions versus small sample sizes (cf. section 4.3). Due to the fact that generally unknown and hidden interrelations between features exist, a backward selection approach in combination with heuristic genetic optimization is chosen to reveal the most important features for VM and also to evaluate all intrinsic information thus avoiding the disadvantage of forward selection in terms of possibly missing individual features with interrelations to other features not yet included in the subset [48], [115].

In addition to the improvement of accuracy and robustness of the prediction models, a major impulse to perform FS is the enormous reduction of data necessary for the processing of VM. Around 20 up to more than 10.000 possibly recorded features in combination with extensive logistical information from different data sources for a single process highlight the demand for smart FS to enable fab-wide VM. Reduction of input data by quality issues during data acquisition and multiplication of input data by consideration of statistical metrics (e. g. mean, median, kurtosis, skewness) or aggregation of features are common examples leading to a wide range of possible feature set sizes. Depending on the consideration of necessary logistical granularities (cf. section 5.2.2), many VM models need to be trained for a single process to achieve highly accurate and reliable predictions. Regarding the conception of a corporate-wide VM system incorporating hundreds of different processes, typical for modern SM, data traffic and storage have to be minimized to avoid inefficient and computational infeasible VM. Expensive data storage in highly available DBs clearly demonstrates the requirement for FS. Also the training of VM models with hundreds or even thousands of instances together with a similar high number of features turns out to fail in MATLAB implementations without special MATLAB intrinsic enhancements due to out of memory errors even for enlarged random access memory of virtual enterprise servers. A linear increase of computation time with the number of features and a quadratic increase with the number of instances can be estimated (cf. section 4.3). Hence, while currently no method is available to encounter all the problems and challenges stated in subsection 4.1.1, the new ERBE algorithm yields the solution.

At first, an assessment of other FS methods (cf. section 3.3) is conducted in order to afterwards compare the performance of some of these FS approaches with the newly invented FS technique.

Thereafter, the SVR kernel is introduced to encounter the challenge of high nonlinearity in data from the simultaneously running physically superpositioned processes (i. e. deposition & sputtering) within the HDP CVD process (cf. section 2.3). At third and fourth the developed FS techniques based on LOO and GA are explained in detail whereas these techniques together build the two elements of the ERBE algorithm. Section 6.5 initially outlines the reasons and advantages of a smart FS algorithm linked up by LOO FS and GA FS followed by the detailed description of the algorithm itself and further complementing considerations.

6.1 Assessment of Feature Selection Methods

Following the actual research in section 4.3, an assessment of appropriate FS methods is performed to find a promising FS technique to ultimately overcome the obstacles of present state of the art and master the challenges described in subsection 4.1.1. In spite of the advantages of computational complexity of embedded methods over wrappers, they are not assessed due to the fact that the interpretability of their FS results is hardly possible for anyone who is not familiar with the intrinsic mathematics of considerable ML methods. The not clearly separable but continuous range of weight vectors (e. g. of SVMs, NN or naive Bayes) rarely differentiate the investigated features or feature subsets. Input variables and especially feature subsets including their interrelations are also hard to interpret in decision trees since splits along the same features are possible in various levels of the tree and feature subsets possibly clustered in the tree are blurred.

Table 6.1 provides important advantages and drawbacks for deterministic and greedy search strategies (e. g. SS), advanced search strategies including filter approaches (e. g. Best First) and metaheuristic search strategies (e. g. Ant Colony Optimization). More detailed characteristics and assumptions as well as descriptions of the advantages and drawbacks of the outlined algorithms are available in the referenced literature (SS: [139]; Backward Elimination: [139]; Hill-Climbing: [139]; Best First: [81], [139]; Branch-and-Bound: [121]; Simulated Annealing: [35], [109], [139]; GA: [117], [144], [139], also cf. subsection 4.3.3; Particle Swarm Optimization: [17], [7], [73]; Ant Colony Optimization: [17], [32]).

Particularly with regard to the stated challenges the favored FS method is required to yield the most important feature to achieve highest accuracy and reliability. SS, Backward Elimination and Hill-Climbing are prone to local optima and by their stepwise methodology do not assess interrelated features or entire feature subsets which is crucial for superpositioned physical processes with strong interaction as in the present use case. The advantage of the Best First search to incorporate efficient filter methods for faster computation at the same time may turn out to be the lack since it may degrade the prediction performance and reliability subsequently assessed by the investigation and comparison of the RELIEF algorithm. The Branch-and-Bound search assesses neither interrelated features nor feature subsets. Characteristics like simple, deterministic and compared to metaheuristics faster computation as common main advantage of the search techniques do not outbalance the existing drawbacks. Simulated Annealing as first metaheuristic search is more robust to local optima but evaluates only a single candidate each

FS Method	Advantages	Drawbacks
Stepwise Selection	simple, deterministic, computational faster than heuristic models	prone to local optima, no assessment of interrelated features, no assessment of feature subsets, risk of overfitting
Backward Elimination	simple, deterministic, computational faster than heuristic models	prone to local optima, no assessment of interrelated features, no assessment of feature subsets, risk of overfitting
Hill-Climbing	simple, deterministic, computational faster than heuristic models	prone to local optima, no assessment of interrelated features, no assessment of feature subsets, risk of overfitting
Best First	simple, deterministic, computational faster than heuristic models, incorporation of efficient filter method	depends on filter method, no assessment of interrelated features, no assessment of feature subsets, risk of overfitting
Branch-and-Bound	simple, deterministic, initial consideration of all feature trees	limited assessment of interrelated features, limited assessment of feature subsets, risk of overfitting, computational intensive with high dimensionality
Simulated Annealing	universal search, global and subspace search, heuristic, robust to local optima	sensitive to algorithm parameters, large set of evaluations, higher risk of overfitting, evaluation of a single candidate each generation, slow final optimization due to random improvements
Genetic Algorithm	universal search, global and subspace search, heuristic, robust to local optima, no influence between candidates	sensitive to algorithm parameters, large set of evaluations, higher risk of overfitting, slow final optimization due to random improvements
Particle Swarm Optimization	universal search, global and subspace search, heuristic, robust to local optima, constant influence between candidates	sensitive to algorithm parameters, large set of evaluations, no solution for non-coordinate systems, higher risk of overfitting, slow final optimization due to random improvements
Ant Colony Optimization	universal search, global and subspace search, heuristic, robust to local optima, constant influence between candidates	sensitive to algorithm parameters, large set of evaluations, problem description to graph matching, higher risk of overfitting, slow final optimization due to random improvements

Table 6.1: Assessment of advantages and drawbacks of FS wrapper techniques.

generation thus gives away the opportunity for a more versatile search at once. GAs like all metaheuristics faces the problems of higher risk to overfitting and sensitivity to dedicated model settings (i. e. adjustments of hyper parameters). In contrast to Particle Swarm and Ant Colony Optimization no influence of candidates within each generation exists for GAs. Particle Swarm Optimization misses a solution for non-coordinate systems and is similar to the deterministic approaches more prone to local optima. At last, Ant Colony Optimization incorporates a constant influence between candidates in each generation but needs to tackle the problem to match the problem description to a graphical representation.

After careful comparison and evaluation of advantages and drawbacks of all FS methods, GA is chosen as one element of the subsequently new FS method ERBE. Robustness to avoid local optima, no influence between candidates in each generation to explore the search space independently and with a wide unbiased variety as well as assessment of interrelated features and feature subsets are the important reasons for this choice.

Many feature projection (e. g. PCA) and dimensionality reduction approaches (e. g. Backward Elimination) have been investigated with partially noticeable success. Nevertheless, no considerable solution for all problems and especially for the difficult and ambitious challenge of knowledge discovery is presented with any of these techniques so far. In addition to many different FS approaches already investigated in chapter 4, a Best First search is conducted including the approved RELIEF algorithm as filter method (cf. section 4.3) with SVR as regression technique. A direct comparison with the new ERBE algorithm is provided in chapters 8 and 9. Thus, the new ERBE algorithm is not only compared to the current state of the art but also to another established technique (i. e. RELIEF & SVR as Best First search) in the area of FS.

6.2 Support Vector Regression Kernel to deal with Nonlinearity

Due to already approved and well-established functionality and good results obtained so far a Gaussian RBF kernel [151] is used for SVR in the present investigation:

$$k(x_i, x_j) := \exp(-\gamma \left\| x_i - x_j \right\|^2) \tag{6.1}$$

The hyper parameter γ within the kernel defines the width of the Gaussian distribution of the RBF. Using the Gaussian RBF kernel for substitution in equation (3.10), the dual optimization problem with nonlinear kernel can be formulated as following:

$$\begin{array}{ll} \text{m}\mathit{aximize} & \left\{ -\frac{1}{2} \sum\limits_{i,j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) k(x_i, x_j) - \varepsilon \sum\limits_{i=1}^{l} (\alpha_i + \alpha_i^*) + \sum\limits_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \right. \\ \mathit{subject\ to} & \left\{ \begin{array}{ll} (i) & \sum\limits_{i=1}^{*} (\alpha_i - \alpha_i^*) = 0 \\ (ii) & \alpha_i, \alpha_i^* \in [0, C] \end{array} \right. \end{array} \tag{6.2}$$

Finally, SVR estimates the best fitting function depending on the three hyper parameters ε, C, γ, which can be optimized by applying a grid search to the SVR ML algorithm whereas the grid search performs a multidimensional optimization for a given range for a specified number of variables. Independency of GA FS of the selected kernel function [160] enables the choice of the well-approved RBF kernel for all parts of the entire new ERBE algorithm (cf. section 6.5).

6.3 Leave-One-Out Support Vector Regression for Feature Selection

In the scope of this thesis no research on mutual information FS was performed due to the superior approach of RFE which was corroborated by recent results [47]. The approach of SVM-RFE performed comparable with an investigated two-layer cutting plane algorithm [110] demonstrating boundaries of SVM-RFE. However, as first novel concept in the present thesis this RFE approach is modified by substituting feature evaluation with introduced LOO assessment of each feature for feature evaluation with intrinsic SVR weights. Upper bounds on the LOO error have been shown and corroborate the investigation of LOO SVR [173].

Algorithm 2: Leave-One-Out Support Vector Regression for Feature Selection

```
Data: Input $X$, Target $Y$
Initialize:
Perform grid search to optimize $C$ & $\gamma$;
forall the features $f_i$ $i \in \{1, \dots, Number\ of\ features\}$ do
    Exclude $f_i$ from $X$;
    Train SVR model on $X \setminus \{f_i\}$;
    Evaluate SVR model by calculating RMSE on independent validation dataset;
    Add RMSE for specific $f_i$ to feature ranking list;
end
Rank by ascending order;
Result: Ranked features by individual impact
```

In terms of computational advantages of SVM-RFE (cf. subsection 4.3.2) only the kernel matrix H has to be recomputed if the weight coefficients α stay constant [49]. However, this assumption cannot be made for highly complex processes as HDP CVD in SM since features present in feature subsets are known to affect each other. The interrelations among the process parameters (i. e. features) are investigated by process experts and important features are controlled but still due to the immense complexity of manufacturing processes not all possible influences are predictable. Thus, as it is not known beforehand whether the impact of each feature stays constant or changes in presence or absence of other features, a new SVR training needs to be performed to evaluate the new feature subset and to recompute α. Therefore and in contrast to SVM-RFE where the weight vectors are computed for each feature individually before a ranking criterion was assessed, this new LOO approach has been developed where the weight coefficients α_i ($i \in \{1, \dots, Number\ of\ Instances\}$) are computed for all remaining features of the feature subset thus including all available information.

One core element of the new ERBE method is the enhancement of SVR RFE with LOO FS which is introduced in algorithm 2. At first, a quadratic grid search is performed to optimize the SVR hyper parameters C and γ around 1 and the inverse of the number of features, respectively, (2). ε is kept constant in scope of the present thesis since it is directly linked to the tolerated deviation of the prediction target and set to a reasonable small error which is not relevant since it is in the range of the possible accuracy of the metrology equipment resulting a final RMSE of ~ 0.4 % (cf. subsection 2.3.3). In LOO one feature is excluded from the feature set (4) and the SVR learning algorithm is trained on the remaining feature subset (5). Afterwards, the impact of this feature is evaluated by the calculated RMSE (6) and added to the feature ranking list (7). Steps (4) to (7) are conducted for all features f_i ($i \in \{1, \ldots, Number\ of\ Features\}$) so that each feature was excluded once. Subsequently, all features are ranked in ascending order (9) as the lack of the single independent most important features degrades the prediction performance the most and by thus yields the highest RMSE and is located at the end of the list. In case of features are only meaningful as combination with others the effect may decrease which is also expected for complex physical processes. Excluded single independent dispensable features indicate a SVR model improvement and a reduction of the RMSE due to less disturbance by introduced noise. These features yield a lower RMSE and are located at the beginning of the list. But this effect may also be softened by feature combinations. However, in order to prevent exhaustive search the feature ranking by individual impact yields a smart tradeoff to distinguish dispensable and crucial features as the output.

Now, a single feature or an entire feature subset can be removed from the initial feature set. As investigated and stated earlier without trading accuracy for speed it is reasonable to eliminate chunks of features at the beginning of the FS process and refine to feature-wise reduction later on (cf. subsection 4.3.2, [49], [115]).

6.4 Genetic Algorithm for Feature Selection

The basic GA algorithm introduced in section 3.5 has been modified for the new concept of smart FS to efficiently consider and extract unknown interrelated features as feature subsets from the entire feature set. Various GA implementations to find essential features by FS are outlined in subsection 4.3.3. According to the emphasized benefits of GAs, the incorporation of GAs empowers FS to assess two important evaluation criteria at the same time namely the prediction performance as accuracy and reliability as well as the optimization of feature subsets. Since the previously introduced LOO FS concept defines a good feature ranking criterion which is not necessarily a good feature subset ranking criterion [49], the by GA enabled optimization for feature subsets is of special importance.

In general, the GA approach aims to optimize an independent and as wide as possible exploration of the feature search space to detect the global optimum or the best local optima which is especially intended in SM to achieve highest possible accuracy while at the same time revealing crucial and maybe unexpected features or feature subsets. Due to the fact that within a chamber all physical process parameters (e. g. temperature, pressure, voltage) have a more or less measurable influences on any other feature (cf. section 2.3), single local optima peaks caused by only one crucial feature subset independent from all other features are extremely implausible. Thus, highly interrelated physical process parameters are expected to be grouped into subsets and various subsets could form bigger feature subset clusters. Therefore, where these subset clusters growing from a wide plateau of many interacting features are expected to form the global optimum or few local optima, continuous improvements in optimizing these few local feature subset clusters are desirable and focused. So, the crossover operation has been skipped in this new FS concept because inversion of a substantial part of the interrelated feature set is not intended in each new generation. Hence, for a sufficient high number of generations with independently (as characteristic of GAs) mutated individuals (i. e. feature sets), the risk to obtain desultory feature subset compositions significantly decreases.

Algorithm 3: Genetic Algorithm with Support Vector Regression for Feature Selection

```
   Data: Input $X$, Target $Y$
 1 Initialize:
 2 Perform grid search to optimize $C$ & $\gamma$;
 3 Set $nG$ := Number of generations to evaluate during GA cycle;
 4 Set $nI$ := Number of individuals to populate each generation;
 5 Set $mR$ := Mutation rate as percentage of total number of features to flip in each GA cycle;
 6 Define RMSE as fitness function for SVR;
 7 Encode feature set as chromosome;
 8 Populate $1^{st}$ generation from initial chromosome;
 9 forall the Generations $G_k$, $k \in \{1, \ldots, nG\}$ do
10     forall the Individuals $I_l \in G_k$, $l \in \{1, \ldots, nI\}$ do
11         Inherit: best individual from parent generation;
12         Mutate $mR$ of all genes;
13         Correct if too few/many genes were mutated;
14         Evaluate fitness function of actual individual;
15         Select best individual if the fitness function was improved;
16     end
17 end
   Result: Ranked features by contribution to feature set impact
```

Algorithm 3 visualizes the sequence of the modified GA for FS which can be adopted for the final ERBE algorithm described in the subsequent section. As for LOO FS, prior any GA computations a quadratic grid search is executed to optimize C and γ (2). A first initialization defines the number of generations (3) together with the number of individuals (4) for population of each generation whereupon the product of both provides the total number of performed SVR trainings and predictions and thus ultimately determining the total computational effort to be spent during the actual GA cycle. Also the mutation rate (5) is set as percentage of the total number of features to be flipped in each GA cycle. Subsequently, the RMSE calculated from real target and prediction is defined as fitness function (6) and the current feature set is encoded as the chromosome (i. e. bit string), (7). The first generation is populated from the initial chromosome at the beginning (8). For all generations (9), the actual individuals of a population (10) are inherited from the best individual of the parent generation (11). Each individual is then independently mutated according to the mutation rate (12) and thus a defined number of genes is activated or deactivated. For the common case that not enough or too much genes are activated/deactivated to meet the defined number of features to be eliminated within each GA cycle, a correction step randomly reactivates/deactivates some genes to achieve the desired number of active genes to evaluate (13) before the fitness function is calculated (14). If individuals of the actual generation achieve a better score with regard to the fitness function than the parent individual they were inherited from, they are selected as candidates for the next generation whereat the best candidate is chosen as parent for the next generation (15). Otherwise, the parents stay for the next generation and the individuals of the actual generation are discarded. At the end of the entire algorithm, one ideal individual evolved with regard to the fitness function whereas the possibly evolving superior feature subsets of all optimized individuals are also of high interest.

The adaptation of a GA for FS enables the concurrent optimization of interrelated features most meaningful only in feature subsets and the elimination of least important features. In addition to the individual feature ranking and optimization of LOO FS, FS by GA complements the new concept of ERBE to find an optimal selection of features by consideration and incorporation of feature subsets. Finally, the entire feature space is reduced by eliminating adjustable chunks of features without forfeiting accuracy for increasing speed as investigated in earlier work [49].

6.5 Evolutionary Repetitive Backward Elimination

6.5.1 Linking Evolutionary GA with LOO Feature Selection

State of the art research is currently not able to apply feature ranking, sequential FS or feature subset selection to achieve efficiency, scalability, knowledge discovery and highest accuracy as essential requirements in SM (cf. subsection 4.1.1). As it is already emphasized and cited earlier, neither pure individual feature ranking or selection nor isolated feature subset selection is suitable to fulfill these ambitious goals in SM. The new ERBE FS algorithm (cf. algorithm 4) primarily published in [94] is developed to leverage the full potential and benefits of both approaches to reveal only the really crucial features (i. e. relevant process parameters) to achieve highest prediction performance in an interdisciplinary and complex environment with superpositioned physical processes in SM. A smart composition of LOO FS and GA FS combines the merits of individual FS and feature subset optimization to finally meet the ambitious objectives in SM. Even more, the new smart ERBE FS concept is very flexible and allows an arbitrary configuration of LOO FS and GA FS depending on the presented use case with underlying complexity.

The new ERBE algorithm (see algorithm 4 below) is organized into three parts I-III to implement the highlighted advantages of:

1. **Part I** Fast initial reduction of noisy information from the feature space by LOO FS
2. **Part II** Feature subset optimization while further reducing the feature space by GA FS
3. **Part III** Fine tuning feature optimization and extraction of crucial features by LOO FS

In order to consider the varying amount of features ranging from 50 to possibly 10.000, each part is subdivided into an adjustable amount of independently executed stages defined by the reduction rate rR which describes the percentage of features to be eliminated in the current stage from the entire or remaining feature set. The introduction of an adjustable reduction rate enables the possibility for advanced control of the computational effort and the entire process of FS by the new ERBE FS algorithm. More stages during any part enable more detailed optimization whereas fewer stages provide the chance to quickly eliminate bigger chunks of features. The stages can be adjusted in a linear manner thus reducing the same percentage of features each stage or in a weighted or any other manner to tune the FS at any time (e. g. more stages at the end for 'high-end' optimization). Artificial features introduced below (cf. subsection 3.3.4) are added to the original feature set for comparison to measure the level of noisy features contained in the feature set. With ongoing feature reduction these artificial features are clearly separated from the crucial features containing most information to achieve high prediction performance with VM. In fact, adding artificial features also adds some variation to the otherwise deterministic LOO FS algorithm. Therefore, 10 optimization cycles are executed within each stage of LOO FS to generate more confidence which features are to be deleted from the feature set while the GA FS is conducted 25 times for more statistical confidence due to its heuristic nature.

Part I focuses on efficient execution of the entire FS and optimization to minimize the necessary computational effort. Especially for larger feature sets with thousands of features fast elimination of chunks of features at the beginning is inevitable since the computational effort of the time determining SVR step is growing linearly with the total amount of features. Thus, if so many features are considered and partially populated (e. g. by statistical moments) part I of the ERBE algorithm maximizes the computational performance by minimizing the computed SVR steps with the new LOO FS concept still incorporating feature dependencies. Artificial features (see below) are added to the original feature set to recognize when a transition from fast initial reduction of noisy information from the original feature space via LOO FS to feature subset optimization via GA FS is required. Where several scenarios are possible to switch over to part II (e. g. after deleting a defined percentage of features) the actual implementation moves on if these artificial features are clearly separated from the remaining real features.

In part II, the incorporation of GA FS is inevitable to optimize entire feature subsets of highly interrelated features with the burden of higher computational effort since many GA generations have to be populated to leverage the advantage of random mutation as statistical heuristic optimization. Thus, the efficiency of the entire ERBE FS process would significantly decrease if the optimization of feature subsets via GA FS would be conducted from the beginning. So, the previously outlined LOO FS is computed initially to quickly reduce the original feature set. While LOO FS is a deterministic approach not capable to perform a global and multisided subspace search, the GA FS enables this functionality by its heuristic nature and possible random walk. Even more, GA FS ensures a high robustness to inferior local optima especially if the elitism of the individuals as feature subsets is reared in several consecutive ERBE stages.

Part III is designed to finally optimize the remaining features containing condensed and optimized feature subsets of interrelated features obtained from the previous GA FS concept as well as individually and independently contributing features. Nevertheless, on the one hand the advantage of GAs heuristic optimization to consider important feature subsets may turn over to more imprecise and rough FS in the end if the mutation rate is not decreased. On the other hand a relative (e. g. percental) decreasing mutation rate might be too small to evaluate all features equally within a preferably small number of generations. Thus, the more fine tuning of feature subsets is conducted by GA FS for final optimization the less efficient the feature space search is performed. The possible case of still existing mainly single irrelevant features after part II would lead to an unnecessary huge amount of randomly populated individuals until this feature is deactivated often enough to achieve sufficient statistical significance if it is evaluated and excluded by itself at all. Therefore, features not only but especially with negligible individual and independent contribution can be differentiated and eliminated faster from the remaining feature subset by conducting the already introduced new LOO FS technique which perfectly solves this problem since it evaluates all features while incorporating feature dependencies. Hence, final feature elimination by LOO FS efficiently further optimizes the already quite good feature subset by quickly reaching a global or local optimum.

For comparison with noisy variables comprising few or no information the artificial features are introduced to distinguish between the three ERBE parts and thus reveal the transition from fast initial reduction via LOO FS to feature subset optimization via GA FS and finally to fine tuning feature optimization again via LOO FS. Since the feature subset and with it the prediction performance constantly improves up to the optimum of the model complexity the conditions for transition to the next ERBE stage are also adjusted. The following list outlines the transition criteria from part I to part II and finally to part III:

1. Part I: LOO FS → Part II: GA FS
   1.1. The first differentiation to real features is achieved if the ratio of the sum of the counts of the selected artificial variables within the 10 % least important features compared to the sum of the counts of the selected real features within the 10 % least important features is ≥ 0.8.
   1.2. The second differentiation to real features is achieved if at least all but one artificial features are within the 15 % of the least important features to prevent very few but highly frequently selected artificial features to cause the transition to an earlier computational intensive feature subset optimization performed by GA FS.
2. Part II: GA FS → Part III: LOO FS
   2.1. The ratio as described in criterion 1.1 is required to be ≥ 1 to achieve an even higher differentiation to real features.
   2.2. In extension to criterion 1.2, the artificial features have to be grouped together without gap as the least important features if at the beginning of part II these features are already grouped as least important ones.

Figure 6.1 visualizes the flowchart of the entire ERBE algorithm with part I designed for fast noise reduction (burnt orange), part II enabling feature subset optimization (lemon yellow) and

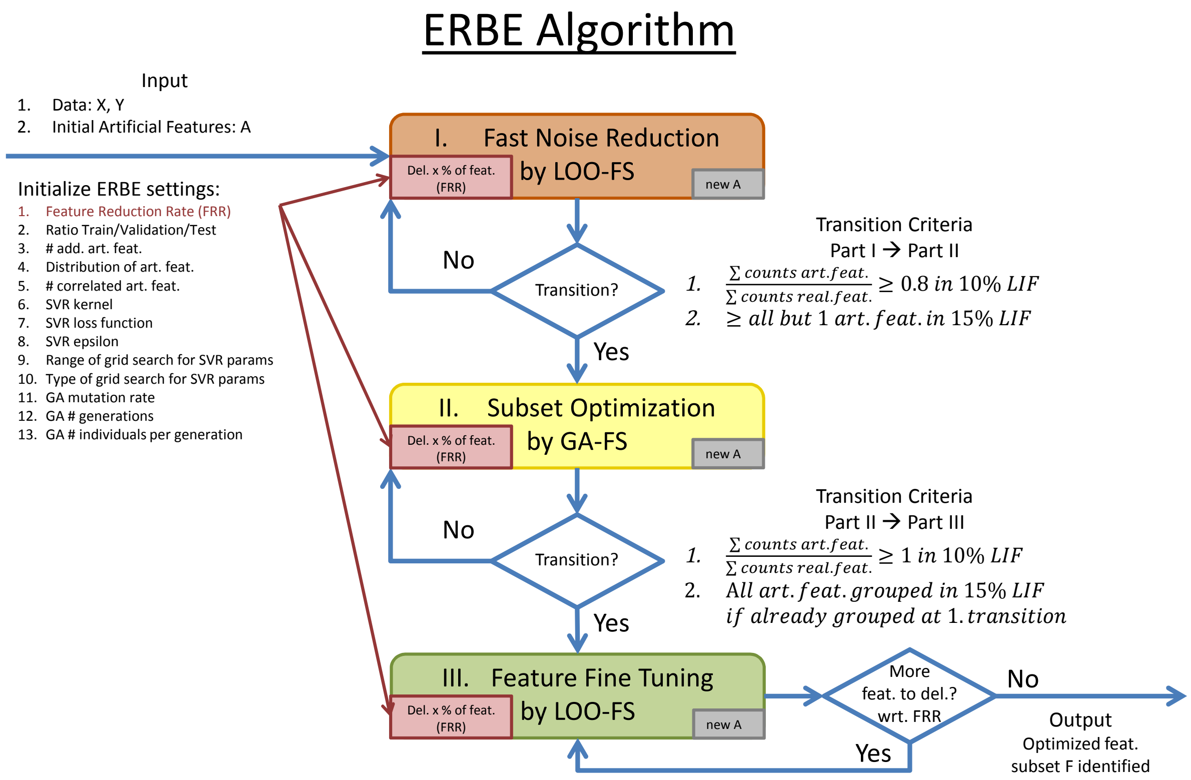

ERBE Algorithm
Input
1. Data: X, Y
2. Initial Artificial Features: A
Initialize ERBE settings:
1. Feature Reduction Rate (FRR)
2. Ratio Train/Validation/Test
3. # add. art. feat.
4. Distribution of art. feat.
5. # correlated art. feat.
6. SVR kernel
7. SVR loss function
8. SVR epsilon
9. Range of grid search for SVR params
10. Type of grid search for SVR params
11. GA mutation rate
12. GA # generations
13. GA # individuals per generation
I. Fast Noise Reduction by LOO-FS
Del. x % of feat. (FRR)
new A
No
Transition?
Yes
Transition Criteria Part I → Part II
1. $\frac{\sum counts\ art.feat.}{\sum counts\ real.feat.} \geq 0.8\ in\ 10\%\ LIF$
2. $\geq all\ but\ 1\ art.feat.in\ 15\%\ LIF$
II. Subset Optimization by GA-FS
Del. x % of feat. (FRR)
new A
No
Transition?
Yes
Transition Criteria Part II → Part III
1. $\frac{\sum counts\ art.feat.}{\sum counts\ real.feat.} \geq 1\ in\ 10\%\ LIF$
2. $All\ art.feat.grouped\ in\ 15\%\ LIF$
$if\ already\ grouped\ at\ 1.transition$
III. Feature Fine Tuning by LOO-FS
Del. x % of feat. (FRR)
new A
More feat. to del.? wrt. FRR
No
Yes
Output
Optimized feat. subset F identified

Figure 6.1: Flowchart of the ERBE algorithm indicating part I - III including the transition criteria w. r. t. the least important features (i. e. LIF) to move on between these parts. Input data X, Y and generated artificial features A are provided with the initial ERBE settings to compute the optimized feature subset F as final output.

part III fine tuning the remaining feature subset (avocado green). All instances with features X and target values Y as well as the initially generated artificial values A are input data for the ERBE algorithm. The artificial features are newly generated in each cycle in every stage of all three parts as represented by a grey box at the lower right side of each part. Prior to the execution the various ERBE algorithm parameters need to be set. These settings include definitions to adjust GA- and SVR-calculations (e. g. GA number of generations & SVR loss function), to introduce artificial features (e. g. Poisson-Distributed - i. e. similar to real features) and to reduce the feature set. The latter named feature reduction rate defines the amount of features to be deleted during each stage in each part illustrated by the red box at the lower left side of each part. For example a feature reduction rate of 10 % would force the algorithm to terminate after 9 stages with 10 % of all initial features remaining in the final subset. Hence, different weights or functions can be used to adjust the feature reduction rate (e. g. higher initial reduction accelerating the algorithm for huge feature sets and smaller reduction emphasizing feature subset fine tuning in the end). The previously outlined transition criteria are listed to demonstrate the threshold for the transition from part I to part II and further to part III. If no more features can be deleted with respect to the initiated feature reduction rate the optimized feature subset F is identified as output.

In the sum, for the given reasons and characteristics the combination of LOO FS and GA FS to form the new smart ERBE FS technique to enable efficient and corporate-wide VM with highest prediction performance adds far more value than the pure concatenation and application of the two approaches themselves. The innovation of a dynamic FS algorithm fuses various advantages of FS and ML to exploit the synergy of LOO FS and GA FS approaches with crafty incorporation of artificial features into the transition criteria.

The ERBE algorithm is designed to always be executed until no more features are left to be further removed even though degradation of prediction performance becomes apparent. In fact, according to the bias-variance tradeoff (cf. section 3.3.4) it is expected to reveal the optimum between model complexity (i. e. size of feature subset) and prediction error. While efficiency is a key factor and highlighted several times five important and independent aspects are distinguished in terms of efficiency:

1. **Fast Feature Selection**: At first, the new ERBE method maximizes the efficiency of the computational effort to conduct FS.
2. **Accuracy**: At second, smart FS ensures highest accuracy and reliability as required for efficient VM (cf. Accuracy).
3. **Scalabillity**: At third, the efficiency of the entire enterprise system is optimized since less required features to run VM scale down continuous productive computational effort of VM, necessary data traffic and expensive DB storage (cf. Scalability).
4. **Efficiency**: At fourth, smart FS as first fully automated FS technique enables the SM industry to efficiently develop and implement fab-wide but still equipment specific VM rollout for all suitable process areas within the entire company (cf. Efficiency).
5. **Knowledge Discovery**: At fifth very important aspect, less expensive experiments are required for research and development (e. g. at the department of unit process development) which can be improved and partially performed more efficiently since process optimization is achieved via knowledge discovery by smart FS (cf. Knowledge Discovery).

6.5.2 The ERBE Algorithm

In more detail describing the new ERBE algorithm (cf. algorithm 4), parts I-III are linked and executed with a variable number of stages in each part depending on the transition criteria. As mentioned above, in order to generate more statistical confidence due to included variations by artificially populated features, for each stage 10 cycles are conducted to evaluate the optimized feature subset of the current stage and to discard the predefined amount of the worst features over all cycles. Thus, for each stage the ERBE algorithm finally yields a series of calculated RMSE values whereof a model complexity curve is created to assess and visualize the complexity of the entire system including the revelation of the most important features (cf. section 4.3, section 8.2).

First of all, all input data as possible matrix containing physical equipment parameters (e. g. pressure, current, gas flow) and logistical information (e. g. basic type, recipe, product) as well as the prediction target as vector are given by X and Y, respectively. During initialization a reduction rate rR (e. g. 10 %) is defined (2) meaning that in each stage chunks of features of 10 % of all initial features are removed. Thus, after all 9 stages 90 % of all original features are eliminated and the expected best 10 % features are left. Basically, rR can be initialized and refined differently for each equipment and process area according to the specific use case with its specific conditions. Furthermore, in order to enable a meaningful evolution of generations in part II, a sufficient number of generations has to be populated to allow the evolution of individuals even after some generations without significant improvement. Thus, the number of generations nG was set to 25 (3) for each stage with the number of individuals nI in each generation set to 5 (4). As next initialization step the mutation rate mR is set to 10 as percentage of the total number of features which are flipped of the current feature subset within each GA FS stage (5). So, 10 % of the remaining features are randomly inverted in each stage of part II. At last, several artificial features based on various distributions (e. g. Gaussian) are defined and created (6) for comparison with the real features to estimate the transition from part I to part II.

Algorithm 4: Evolutionary Repetitive Backward Elimination Feature Selection

```
   Data: Input $X$, Target $Y$
1  Initialize:
2      Set $rR$ := Reduction rate as percentage of features to reduce during each stage;
3      Set $nG$ := Number of generations to evaluate during GA cycle;
4      Set $nI$ := Number of individuals to populate each generation;
5      Set $mR$ := Mutation rate as percentage of total number of features to flip;
6      Create artificial features for comparison with real ones;
7  Part I: Fast elimination of features contributing mainly noise compared to artificial ones
8  repeat
9      Perform LOO FS;
10 until switch to feature subset optimization is observable by artificial feature comparison;
11 Part II: Feature subset optimization incorporating crucial interdependencies
12 repeat
13     Perform GA FS;
14 until switch to fine tuning feature optimization is visible by artificial feature differentiation;
15 Part III: Fine tuning feature optimization up to most important $rR$ % of features
16 repeat
17     Perform LOO FS;
18 until remaining reduction to final $rR$ is achieved;
   Result: Optimized feature subset for highest prediction accuracy and reliability
```

After initialization of the required ERBE algorithm hyper parameters, part I (7–10) is executed to quickly perform an initial reduction of noisy features ensuring an efficient execution of the ERBE algorithm until a transition from fast initial reduction via repetitive LOO FS (9) to feature subset optimization via repetitive GA FS is achieved by comparison with artificial features (10). In each of the repetitive stages of part I, the LOO FS (cf. algorithm 2) result is compared to the introduced artificial features. If these artificial features are separated from the residuary features with regard to the transition criteria the new ERBE method moves on and part II (11–14) is conducted subsequently. Here, in each of the repetitive stages GA FS (cf. algorithm 3) is performed (13) to emphasize feature subset optimization while incorporating crucial feature interdependencies based on physical interrelations. The artificial features are kept in the DS to constantly obtain an assessment of the optimized feature subsets compared to noisy and artificial information where clear differentiation between artificial and real features causes the move to part III (15–18) again with regard to the transition criteria. Following, part III performs a fine tuning feature optimization in repetitive LOO FS stages again (17) and is linked to part II to obtain an improved differentiation of moderate to most important features. Due to the fact that LOO FS is computational more efficient than GA FS (cf. subsection 6.5.3) and crucial feature interdependencies are already considerably incorporated, faster feature optimization up to the most important $rR\,\%$ (e. g. 10 %) (18) of features is achieved in part III still including artificial features for comparison. Finally, the optimized feature subset for highest prediction accuracy and reliability is obtained as result. In a generic approach and independent of any equipment or process areas, the result of the new ERBE FS yields the finally optimized feature subset for highest prediction accuracy and reliability in VM.

6.5.3 ERBE Considerations

In order to demonstrate the generic approach of smart FS, two different process equipment (i. e. AMAT Centura & AMAT Producer, cf. section 7.3) are investigated in the present work still containing 80 & 198 features after rigorous DP. An approach with less strict DP filling up features containing NaN values (e. g. with the mean of the available values), more activated sensors and more considered statistical moments (e. g. kurtosis, skewness) for the available features will easily produce bigger datasets up to several thousands of features. However, the computational effort to run ERBE FS can easily be approximated due to linear increase feature-wise compared to quadratic increase instance-wise while the reduction ability can be assured due to the detectable strong correlation of several statistical moments of populated features as in earlier observed investigations (cf. chapter 4).

A significant number of features not containing useful information can be deleted quite early during execution of the algorithm. Due to the fact that the total number of computational intensive SVR evaluation steps for each FS stage is smaller for LOO FS compared to GA FS, the former is predestinated for part I to rapidly reduce the entire feature space. The total number of SVR evaluation steps to be conducted by GA FS and LOO FS is calculated in equation 6.3, respectively. The example is provided for the total number of 80 features (i. e. already including the five artificial ones) for ERBE FS on the DS of AMAT Centura.

$$\begin{aligned} \#SVR\text{–}Eval_{GA} &= nG * nI = 25 * 5 = 125 \\ \#SVR\text{–}Eval_{LOO} &= \#features = 80 \end{aligned} \tag{6.3}$$

In order to achieve a statistical significant assessment of the heuristic GA FS technique, the total number of evaluations as product of the number of generations and individuals naturally needs to be linked and adjusted with the total number of features. Thus, for feature sets with more than thousand features a meaningful total number of GA FS evaluations could easily exceed 2000 with 20 individuals in 100 generations compared to constantly 1000 LOO FS evaluations. Trivially, in a heuristic approach the total number of evaluations clearly needs to exceed the total number of features.

The repetitive design of the ERBE method to delete chunks of features at the beginning is corroborated by the fact that there exist only significant differences for medium to smaller feature subsets whereas in early optimization phases the removal of chunks is not affecting the prediction accuracy [49], [115]. A more detailed investigation of a feature-wise reduction at the end of a FS approach is also motivated in the work of [49] approving the new LOO FS concept for higher feature tuning after GA FS again.

In order to avoid model optimization on a specific time period in the available DS for the ERBE algorithm, to include rare adjustments of equipment settings in later instances and to increase the prediction accuracy of SVR by incorporating more characteristics over a longer time period (e. g. drifts), the chronological order is randomized by shuffling of the instances of the entire DS. The investigation of smart FS does not simulate the productive VM environment where the test DS is chronologically separated to assume future data (cf. section 7.2). The new ERBE algorithm intends to reveal the most important features so it is reasonable to include and shuffle data of the entire time period. For each cycle of all ERBE stages as well as for the evaluation of the finally investigated feature subsets data are shuffled independently. Thus, many SVR models are optimized in various ERBE stages on always newly mixed DSs.

Furthermore, a quadratic grid search is implemented to optimize the SVR regularization parameters: cost factor C and kernel distribution width γ. Initially and before each stage, C and γ are optimized for the remaining feature subset and then adopted for all evaluations within this stage where it is reasonably assumed that C and γ serve well for each specific feature subset per stage. An optimization of the SVR regularization parameters on a global feature set does not necessarily guarantee optimal SVR parameters for all possible feature subsets [31], [139]. The MATLAB source implementation of SVR is enhanced by only computing half of the internal H matrix containing the kernel results since H can be mirrored at the diagonal. Also a MATLAB internal quadratic programming is used to solve the interior-point convex optimization algorithm.

According to the relevance of features and instances (cf. section 4.3) and in comparison to real features as already mentioned above, five artificially generated features are introduced into the dataset. These artificial features are created to simulate some variations which are present in available data incorporating various distributions (i. e. Gaussian & Uniform) and correlation. Two features based on a normal distribution ranging from 0 and 1 and one feature based on an uniform distribution also normalized from 0 to 1 are added. In addition, one feature of both distributions was duplicated and doubled to serve for the assessment of highly correlated features without adding any information.

Summary: Following elaborated state of the art in chapter 4 and the description of the required VM system in chapter 5, the present chapter initially conducts an assessment of existing FS methods emphasizing advantages and drawbacks.

A kernel extension of SVR enables the new FS concept to overcome frequent monotonicity assumptions and therefore to deal with nonlinearity as it is available in highly complex superpositioned physical processes common in SM. Corporate-wide deployment of VM and most efficient implementation of VM in SM requires an universal FS approach applicable for all equipment and process areas regardless of linear or nonlinear relationships within data. Hence, the investigated kernel extension completes this generic approach of smart FS.

The new LOO FS and GA FS techniques are investigated as well as developed and implemented for the specific demands of VM for the challenging environment in SM industry. Even more, these sophisticated methods are linked together to ultimately build the efficient but still very adjustable and adaptable smart ERBE FS algorithm to reveal only crucial features absolutely necessary to achieve highest prediction performance and reliability. Part I-III are optimized on each other to leverage the full potential of the advantages of LOO FS and GA FS even though the execution of these parts is very flexible. Thus, these parts and with it the concepts of LOO FS and GA FS can be dynamically linked and multiplied in any order to form an evolutionary and highly dynamic approach consistently incorporating benefits of both techniques. Finally, a flexible amount of features from 20 up to more than 10.000 can be reduced and optimized whereas the introduced stages enable a customization of the method to focus on constant feature reduction, successively refined feature optimization or any other setup.

The extension of the new ERBE FS approach for concurrent introduction of artificial features to emphasize the differentiation of less important features adds one more adjuvant functionality (i. e. especially at the beginning while in the end it is clearly visible that artificial features are constantly performing worse than more important ones).

Chapter 4 extensively outlines the current state of the art of VM and FS to improve VM but still the stated problems and challenges could not be tackled by any investigated FS approach so far. While never before any other concept or algorithm completely met these ambitious objectives in VM, the newly invented ERBE FS algorithm introduced in this dissertation entirely achieves this final goal to ultimately reveal the best feature subset yielding highest prediction performance and reliability. Hence, the ERBE FS algorithm enables a concept for smart FS to improve current VM to become advanced VM by smart FS.

The next chapters specify the conducted experiments and present the achieved results of the new ERBE FS algorithm itself as well as a comparison with the established but so far not evaluated RELIEF filter method and the pure application of the previously highlighted LOO FS approach without subsequent feature subset optimization by additional GA FS.

7 Experiments

With respect to the two key aspects of the present thesis (i. e. the advanced Virtual Metrology system and smart Feature Selection) comprehensive experiments are conducted to assess the attainment of the projected goals essential for the implementation of the developed advanced Virtual Metrology system at Infineon. Initially overall Data Preparation is outlined.

7.1 Overall Data Preparation and Modeling

In the following, the overall applicable DP steps data formatting, dataset compilation, feature translation and feature selection are described. Extensions of dataset compilation and FS as well as instance selection are performed specific for the evaluation of the advanced VM system and the new ERBE FS algorithm and therefore further outlined in the next section.

Data Formatting: For the development of the advanced VM system the logistical parameter – Logistic_1 – is converted from character to numeric format (i. e. $\in \mathbb{N}$) and thus provided as input for later SVR induction learning. All other features are available as floating point numbers and no further formatting was required.

Dataset Compilation: All available data from different sources (cf. section 5.2) are merged to obtain a coherent DS. In terms of supervised learning, all instances without metrology data and necessary logistical information had to be removed from the DS resulting in an enormous reduction of data due to the applied sampling rate containing a comparable small amount of metrology data and partially missing logistical parameters required for bijective data assignment. On the one hand, a sampling rate of 10 % – 15 % measured wafers for a HDP CVD process is quite common in SM whereas on the other hand from those remaining measured wafers some more are removed due to the stated missing logistical parameters finally yielding a reduction of the original data by 91 %.

Feature Selection: During the task of DS compilation all features from metrology and process data containing only NaN values are immediately removed. Missing values within a single feature are replaced by the mean of this feature vector. In the special and rare case of only one remaining value within a single feature thus generating a constant feature vector the feature is marked as redundant and also removed from the feature set. According to figure 7.1, overall FS is performed again after merging metrology and process data to delete instances without measurements.

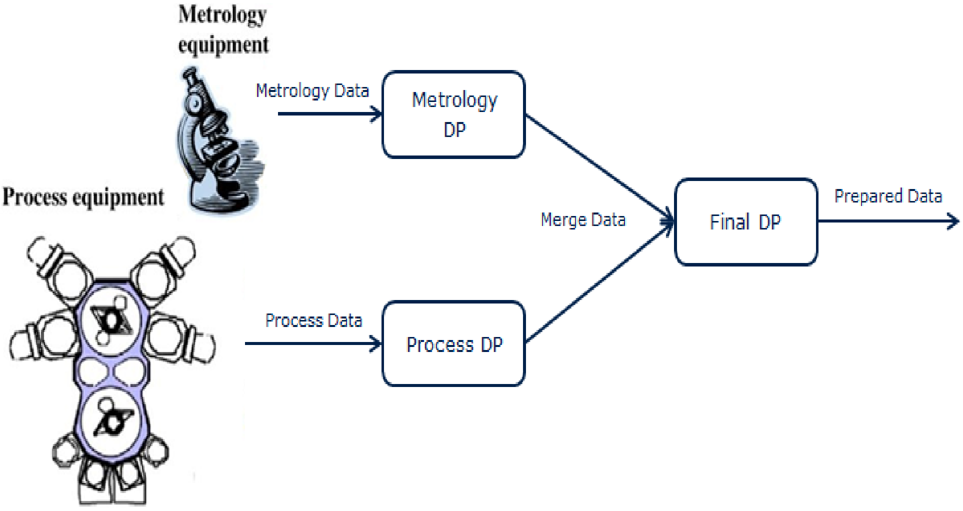


Figure 7.1: Data Preparation Sequence for overall FS.

Feature Translation: In order to avoid domination of features with large unit scales over smaller scaled features, normalization according to equation 7.1 is performed for all features individually to map all values to the interval [0,1] where the borders represent the smallest and highest value in the entire feature vector, respectively.

$$\mathbf{x}_{norm} = \frac{\mathbf{x} - Min(\mathbf{x})}{Max(\mathbf{x}) - Min(\mathbf{x})} \tag{7.1}$$

7.2 Advanced Virtual Metrology System

The advanced Virtual Metrology system implementation has been outlined in subsection 5.2.2 whereas the experimental setup to validate the new system is given below. After outlining the specific DP for this challenge, the performed VM prediction on new data as a comparison of the SVR models trained and optimized on a complete dataset versus a MW dataset approach is decribed.

7.2.1 Data Preparation

The DP tasks data formatting and feature translation are already described above and only the additional DP task instance selection together with the extension of dataset compilation and feature selection specific for the advanced VM system are highlighted below.

Instance Selection: All instances associated with obvious outliers (e. g. exceeding the 3σ standard deviation range) exclusive in either any feature or the target are removed from the training and test DSs due to the sensitivity of the SVR performance to inherent outliers. Nevertheless, the total amount of deleted instances (i. e. ≤ 16) is negligible compared to the total amount of instances (≥ 2087).

Dataset Compilation: In addition to the DS covering two years of production initially used for the investigation of the ERBE FS algorithm, various similar expanded DSs of other equipment of the same type within the HDP CVD work center (i. e. AMAT Centura, cf. subsection 2.3.1) are acquired and compiled to be used for SVR model training and validation. Additionally, a disjunct DS for a subsequent half-year time period of production is acquired for each equipment and compiled to be exclusively used for testing of the trained and validated VM model based on the most important features as result revealed by the ERBE algorithm. The logistical granularity is only defined by the equipment, a randomly chosen process chamber of each equipment and the considered process recipe.

In terms of the size of the final training DS, a complete DS containing all instances is compiled as well as a MW DS containing only a fourth of instances defined as the most recent 25 % of the DS. Without performing cross-validation 80 % of the shuffled training data in both DSs are used to train the final SVR model and the remaining 20 % served as validation data. The chronological order of data within the training and validation DS is not required for prediction by SVR whereat the chronological order within the independent test DS is kept to realistically simulate the application of the VM models to accrued future process data.

Feature Selection: The optimized feature subset selected by the newly developed ERBE method (cf. section 6.5 and table 8.7) is used for productive online prediction of the HDP CVD LT.

7.2.2 Virtual Metrology Prediction

For both, the complete and the MW DS, a linear grid search was performed to optimize the SVR model parameters C and γ (cf. section 3.4). As illustrated in equation 7.2 and based on common practice as a rule of thumb, the cost factor C and the width of the Gaussian distribution γ were optimized around 1 and the inverse of the sum of the incorporated features (cf. subsection 8.2.1 – e. g. 0.1), respectively. The steps of the grid search were chosen linearly around C and γ.

$$\begin{gathered} C \in \{10, 7.5, 5, 2.5;\ 1;\ 0.75;\ 0.5;\ 0.25;\ 0.1\} \\ \gamma \in \{1;\ 0.75;\ 0.5;\ 0.25;\ 0.1;\ 0.075;\ 0.05;\ 0.025;\ 0.01\} \end{gathered} \tag{7.2}$$

Virtual Metrology Model Training on complete and MW Dataset: Five equipment are investigated exemplarily for one out of three randomly chosen process chamber and for a single process recipe over a two and a half-year time period lasting from 05/01/2011 to 10/31/2013 in order to assess accuracy and reliability of the predictions delivered by the advanced VM system. Table 7.1 specifies for each equipment EQ_{1-5} the process chamber CH_{1-3}, the number of initial and remaining instances within the DS, the number of removed outliers as well as the number of instances assigned to the complete training, MW training and test DS.

Virtual Metrology Model Training on updated MW Dataset: Due to most recent major adjustments of equipment and/or recipe settings for two of the investigated equipment (EQ_1

Equipment	Chamber	# Instances	# Outliers	# Remaining Instances	# Train Complete	# Train MW	# Test
EQ_1	CH_1	3427	10	3417	2734	685	683
EQ_2	CH_2	2087	10	2077	1662	417	415
EQ_3	CH_2	2350	8	2342	1874	470	468
EQ_4	CH_3	3741	16	3725	2980	746	745
EQ_5	CH_1	2425	8	2417	1934	485	483

Table 7.1: MW (i. e. 25 % of recent data) and complete DS for VM.

& EQ_2) on the verge of the half-year time period reserved for testing and the resulting poor prediction performance (cf. section 8.1), an updated MW DS is used in addition to the previously described DSs to perform a more elaborated analysis. In this sense, the original test DS of the half-year time period is halved whereas the first half is used as new training DS and the second half is reserved as new test DS. For EQ_2, the first 19 instances of the new training DS include data prior to the major equipment adjustments. Thus, these instances are excluded from the training DS and the first 19 instances of the new test DS are used instead explaining the difference of $209 + 187 = 396$ instances to originally 415 instances. The number of original test instances and the sum of training and test instances can differ due to the automatically applied rounding of the indices. Table 7.2 extends the previous table in terms of the updated MW DS.

Equipment	# Test original	# Train updated	# Test updated
EQ_1	683	342	342
EQ_2	415	209	187

Table 7.2: Updated MW DS for VM.

7.3 Smart Feature Selection

The ERBE algorithm is highlighted in detail in section 6.5 whereas the experimental setup to approve its usability for VM in productive SM is outlined in the following. Again, well adapted DP for the ERBE testing are provided first. Furthermore, additional experiments are performed for different production equipment to verify the generic approach of the FS technique with the investigated induction method and its applicability in other process areas (cf. subsection 7.3.2). The SVR model parameter ϵ is set to allow ~ 0.2 % deviation from the predicted target to prevent overfitting which is also reasonable regarding the accuracy of ~ 0.4 % for the physical metrology measurement (cf. subsection 5.2.3). Finally, a comparison of the new ERBE FS, the established RELIEF FS filter and the LOO FS wrapper algorithm is conducted (cf. subsection 7.3.3).

7.3.1 Data Preparation for Feature Selection

Overall applicable data preparation for the assessment of ERBE FS is already outlined in section 7.1 including the comprehensive tasks of data formatting, dataset compilation, feature translation and FS. Thus, only the FS specific DP step of instance selection as well as the extension of dataset compilation and FS are summarized below.

Dataset Compilation for the ERBE, RELIEF and single LOO FS Algorithms

ERBE FS for AMAT Centura: A complete DS covering two years of productive data from process chamber CH_1 of equipment EQ_4 (i. e. 05/01/2011–04/30/2013) is investigated for the development of the ERBE algorithm. Due to the fact that all information shall be included for knowledge discovery, data of the entire DS are shuffled before any separation of DSs is performed. Subsequently, 80 % of all data are used for training (i. e. 60 %) and validation (i. e. 20 %) and the remaining 20 % of all data are separated to compile an independent and unseen test DS. The initially available features of the AMAT Centura production equipment for ERBE FS are listed (cf. appendix A.4.1) including category, unit and a short description according to section 2.3.

ERBE FS for AMAT Producer: In order to demonstrate the generic approach of the new ERBE FS algorithm it is also run and tested for a different manufacturing process (PECVD) on a different production equipment (AMAT Producer) in a different work center (cf. differences of HDP CVD and PECVD in section 2.4) [4], [5], [127], [128]. On this equipment, each of the three installed process twin-chambers contain two identical chucks to process two wafers within the twin-chamber at the same time. Thus, six wafers could be handled concurrently. For the investigation a single process for the deposition of a silicon oxide base layer onto a metal layer stack is selected. Data for a time period of 18 months (i. e. 05/01/2012–10/31/2013) are prepared and purified according to the DP steps described in subsection 7.2.1. From initial 1389 instances for 198 features 826 instances are remaining which in contrast to the FS on the AMAT Centura clearly falls below a general rule of thumb of at least ten instances per feature [51]. For the sake of completeness the features for the AMAT Producer are listed in the appendix on page xxiii.

RELIEF: Due to the fact that RELIEF is a correlation-based filter FS technique, no model is trained during FS and thus no separated DSs for training and validation are needed. Since the ERBE FS and the LOO FS techniques randomly use 80 % of all data for training (60 %) and validation (20 %) and 20 % for final testing, the DS to evaluate the RELIEF algorithm for data of the HDP CVD process manufactured in the equipment AMAT Centura also contains 80 % of all data and 20 % are separated as test DS.

Single LOO FS: Single LOO FS essentially performs only the first stage of the ERBE FS algorithm without further repetitive LOO FS or GA FS improvements. Therefore, the same DS as for ERBE FS for AMAT Centura described above is used for evaluation of LOO FS (i. e. 60 % for training, 20 % for validation & 20 % for testing).

Instance Selection for the ERBE, RELIEF and single LOO FS Algorithms

The total amount of deleted instances due to obvious outliers (i. e. exceeding the 3σ standard deviation range) exclusive in either any feature or the target degrading the prediction performance of SVR is negligible (i. e. 2 out of 2701 instances) finally yielding 2699 instances for the HDP CVD process manufactured on the equipment AMAT Centura.

Feature Selection for the ERBE, RELIEF and single LOO FS Algorithms

Despite the outlined DP, no previous algorithmic FS is previously executed to remain all original features within the entire feature set. Instead, according to section 6.5, five artificial features are added to improve the capability of monitoring the effectiveness in terms of the revelation and elimination of redundant and noisy features.

7.3.2 ERBE Feature Selection

The ERBE FS algorithm has been developed and initially investigated for data of the manufacturing equipment AMAT Centura. In order to confirm the genericity of the ERBE FS approach, the new algorithm is also applied to data of the AMAT Producer production equipment. The experimental setup for both investigations is detailed in the following.

ERBE Feature Selection for AMAT Centura: Eight out of 80 initial features (including the 5 artificial ones) are eliminated during each stage of the ERBE FS algorithm according to the defined reduction rate rR of 10 %. The reduction rate of 10 % is chosen because for a comparably small original feature set with 80 features the transition where artificial features are clearly separable from real features is expected to appear quite early. In contrast, for scenarios with thousands of features an initial rR in decreasing manner (e. g. from 30 % down to 5 %) could be defined to reduce the bigger and by this computational more expensive feature set more quickly hence performing the feature search more efficient. For a better comparison of ERBE FS between AMAT Centura and AMAT Producer the reduction rate is kept constant. As a benefit of the new ERBE FS technique, rR can be flexibly adjusted depending on the size of the original feature set allowing possible filling instead of elimination of sparsely populated features. In order to assess redundant and less relevant features as well as to push the feature subset selection, the artificial features are not removed from the feature subset even in case of performing worst. Thus, the point of differentiating noisy features from important ones shall be clearly revealed and after this point the artificial features yield constantly worse results than features containing crucial information. Around 20 % of the least important features are listed for each stage including a counter of how often a feature is ranked among this portion of least important features during the ten cycles per stage. At the end of each stage, the 10 % features with the highest counts are removed from the feature subset whereas in case of equal counts the features are selected randomly for elimination. The GA FS within ERBE FS optimizes the feature subsets with focus on rearing elitism thus only including the best features of the finally best individual. Hence, the features selected the fewest out of the best individuals of 25 generations (i. e. the individuals yielding an improvement in prediction performance) with five individuals in each generation are removed in all 25 cycles. After all ERBE stages have been conducted, ten SVR models are trained on the remaining feature subset and the accuracy is averaged. A complete list of all ERBE FS parameters and settings is provided in the appendix on page xxvi.

The potential for further optimization of the ERBE algorithm is assessed by manually investigating the final features to find a possibly global optimal feature subset. The ERBE FS result (i. e. final feature subset) is initially compared to an independent Expert Selection (ES) of two process experts of all available features and subsequently manually changed by adding and removing features which were left after the 7^{th} ERBE stage or from the physical background knowledge of the process experts expected to contain important information. Finally, a feature subset is selected as input to assess the VM system and assumed as global optimum which yields very accurate predictions on the one hand and simultaneously reduces the feature size to only a small fraction of the original feature size on the other hand.

ERBE Feature Selection for AMAT Producer: From 203 initial features including the five artificial ones, a reduction rate of 10 % is set and 20–21 features are eliminated during each stage. Again the artificial features are kept in the feature subsets and the procedure of FS is carried out just as for the AMAT Centura.

7.3.3 Comparison of ERBE, RELIEF and single LOO Feature Selection

The RELIEF technique generates a correlation-based feature ranking where all features are included and sorted in descent order starting with the most important feature. In order to achieve a comparable evaluation similar to the ERBE FS technique, ten runs are conducted and the average of all results is considered. Finally, the features most often ranked as most important features are selected and SVR models are trained for feature subsets composed of these selected top 3, 5, 8, 10, 12 & 15 features to predict the target on the independent test DS.

Furthermore, the ERBE FS technique is compared to the LOO FS wrapper approach to also assess the pure improvements achieved by the incorporation of SVR in contrast to the correlation-based RELIEF filter method. The investigation is performed if the pure application of the developed LOO FS is already sufficient to obtain an optimal feature subset and thus allowing to possibly omit further optimization by GA FS. Stage one of the ERBE FS technique performs LOO FS as wrapper method with significantly less computational effort compared to the entire execution of ERBE FS and by means of the again best 3, 5, 8, 10, 12 & 15 ranked features SVR models are trained and tested on the independent test DS.

Summary: After detailed description of the newly invented ERBE FS algorithm in chapter 6 and the by this advanced VM system outlined in chapter 5, specific experiments are designed to obtain comprehensive results of their performance. The new ERBE FS technique is tested for two different processes (HDP CVD & PECVD) manufactured by different equipment (AMAT Centura & AMAT Producer) to evaluate the capability to reveal the most important features thereby optimizing the prediction performance. Furthermore, a clear distinction is drawn between the ERBE FS method and an established filter and wrapper approach (RELIEF FS and LOO FS, respectively). The new advanced VM system is deployed on DSs of five different equipment (AMAT Centura) to assess the challenges and problems of subsection 4.1.1. The upcoming chapters present results and discussions and are also organized in two parts for VM and FS.

8 Results

The results of the experiments described in the previous chapter are presented below to firstly evaluate the implemented advanced Virtual Metrology system for the highly complex High Density Plasma Chemical Vapor Deposition process on various production equipment of the same type and secondly to analyze the new Evolutionary Repetitive Backward Elimination Feature Selection algorithm for production equipment of different types and for different manufacturing processes as well as in comparison with the RELIEF filter and the Leave-One-Out wrapper algorithm. The prediction performance in terms of accuracy is measured by CV(RMSE) (cf. section 3.6.1) and analogous CV(MAE) as percentage of deviation from the target. Reliability is assessed via the R^2 and the number of correctly predicted outliers outside Upper Control Limit and Lower Control Limit for the observed target values in relation to the total number of outliers which is measured by the sensitivity (cf. subsection 3.6.3). The specificity is not considered in the scope of this thesis because highest specificity ($\geq 98\,\%$) is constantly achieved [87] and expected by process engineers since a Virtual Metrology system fails to work efficiently if too many false alarms are raised.

8.1 Advanced Virtual Metrology System

Independent and unseen data of a half-year time period of all five AMAT Centura production equipment are used for real-time tests while the VM models are trained on historical data covering a time period of two years for the complete DS or a couple of months for the MW approach. All VM models are built on the feature subset selected by the ERBE FS algorithm and are further optimized as outlined in subsection 8.2.1.

Tables 8.1 – 8.5 list the measured metrics for all VM models trained on the DSs of all investigated AMAT Centura production equipment EQ_1 – EQ_5. Furthermore, charts of observed target values (green) and VM predictions (red) based on the complete and MW DSs (and if necessary also the updated MW DSs for EQ_1 and EQ_2) are presented. Due to the fact that a productive online environment is simulated, the relative layer thickness as target is plotted against the wafer count only for test data on these charts. UCL and LCL for the measured target are illustrated as horizontal blue lines to highlight the very narrow process specification required for nanoscale designed high-tech products fabricated in SM. For EQ_1 and EQ_2 relevant features are illustrated for enhanced assessment of an updated MW approach where the normalized feature value is plotted against the total number of evaluated instances (i. e. the wafer count) within the combined training and test DS.

8.1.1 Equipment 1

Training DS	CV(RMSE)	CV(MAE)	R^2	Sen.	# Outliers	# Test	# Train
Complete	3.85	3.65	-107.69	0	4	683	2734
MW	4.54	4.32	-149.72	0	4	683	685
updated MW	1.02	0.82	-5.79	0.5	2	342	342

Table 8.1: Results for EQ_1 for complete, MW and updated MW DS: Evaluation of accuracy by CV(RMSE), CV(MAE) and R^2 as well as reliability by sensitivity including the number of outliers and instances within test and training DS.

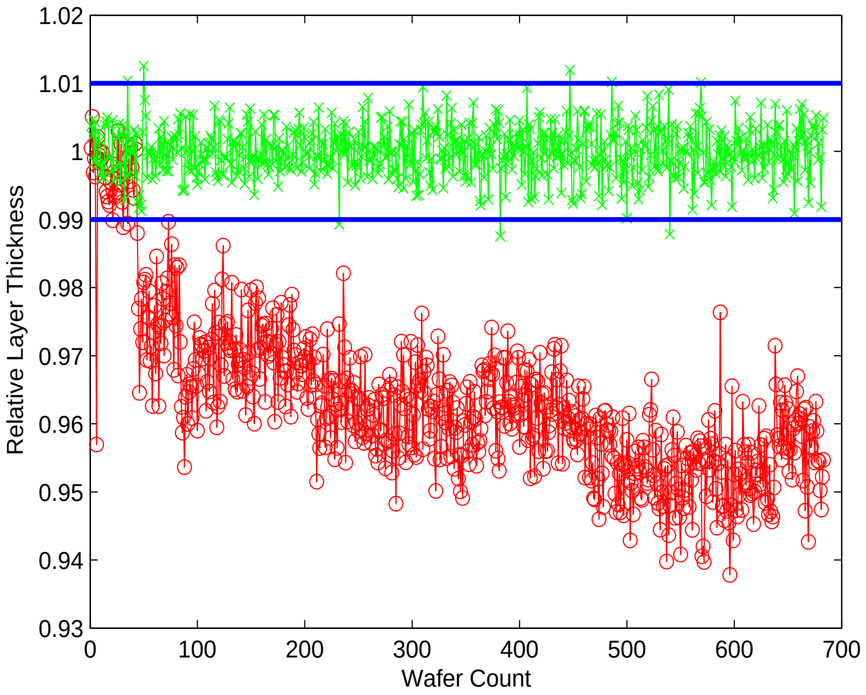


Figure 8.1: For EQ_1 the prediction (red) of the relative layer thickness for all wafers of the independent test DS for the VM model trained on the *complete* DS based on two-years time period shows significant degradation from ~ wafer 50. The measured target (green) is provided with UCL and LCL (blue).

The primary training of the VM prediction model for EQ_1 obviously failed for both used DSs (i. e. complete & MW) identified by all evaluation measures (i. e. RMSE, MAE, R^2 & sensitivity) and impressively visible in the figures 8.1 and 8.2.

Even though the obtained RMSE of 3.85 %–4.54 % is comparable to some results of various publications stated in section 4.2, it is not considerable for implementation of VM since control limits of process specifications for nanoscale designed high-tech products in SM constantly shrink. Feasibility and applicability of VM depends on predictions with an accuracy of < 1 %. Right at the transition from the training to the test DS the equipment was subject to a major maintenance activity described in subsection 2.3.2. The required readjustment of several process parameters

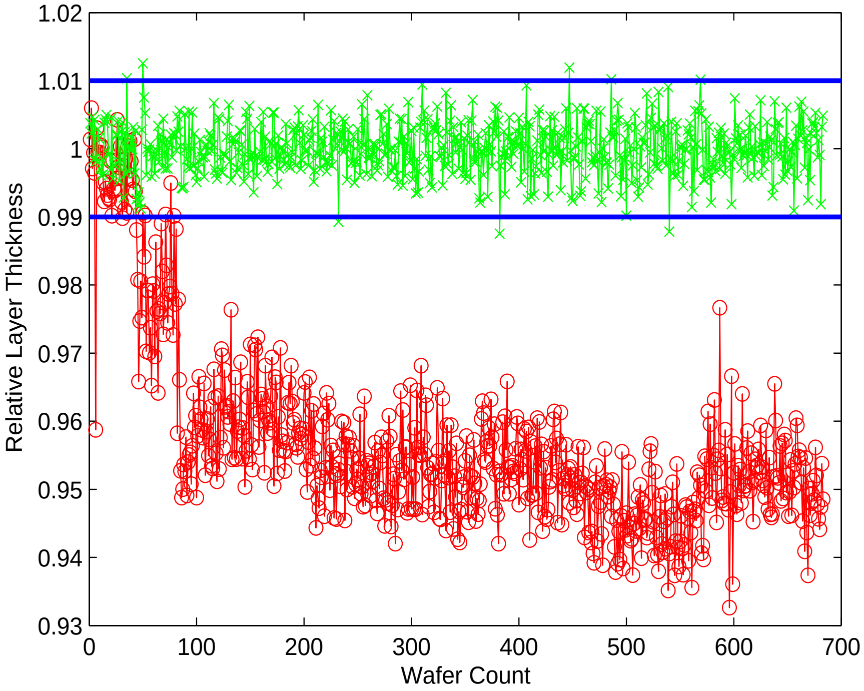


Figure 8.2: For EQ_1 the prediction (red) of the relative layer thickness for all wafers of the independent test DS for the VM model trained on the *MW* DS based on 6 months time period shows significant degradation from ~ wafer 50. The measured target (green) is provided with UCL and LCL (blue).

can be observed in two of the eight most important features selected by the ERBE FS algorithm (cf. table 8.7) as displayed in figures 8.3 and 8.4 for the last ~650 instances.

Figure 8.3 illustrates the normalized temperature of the input feature Temperature_Dome_3 for the combined training and test DS. Clearly visible is the offset in the feature values just before and after the periodical maintenance at about wafer number 2800. Similarly, figure 8.4 displays the shift of Voltage_Chuck_1 to an so far unknown range of feature values 0.65 to 0.8. For comparison figure 8.5 shows Power_TS_1 as one of the crucial features with typical variation within the DS not affected by maintenance activity. The significant change of two of the most important input features causes the VM model to perceptibly degrade and to fail to accurately and reliably predict the relative layer thickness.

The prediction conducted for the updated MW DS (cf. subsection 7.2.2) yields improved results illustrated in figure 8.6 especially for the first third of wafers in the test DS. Nevertheless, regarding the significant improvement of all evaluation measures for the MW DS (cf. table 8.1), the result for EQ_1 indicates a possible solution for achieving acceptable prediction performance after maintenance intervention even though further investigations are necessary to develop a more effective MW approach.

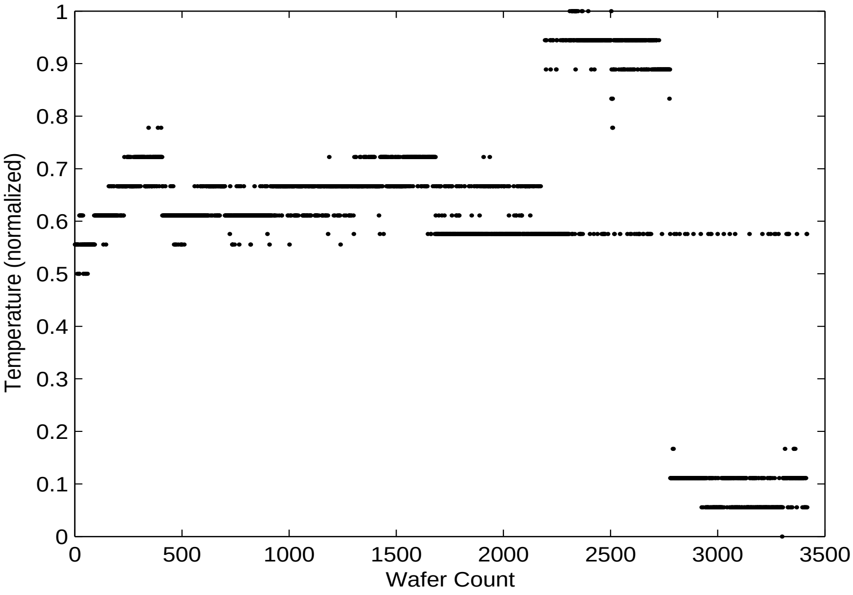


Figure 8.3: Due to readjustment of several process parameters for EQ_1 related to a periodical maintenance, Temperature_Dome_3 as one of the crucial input features selected by ERBE FS shows a significant shift to a so far unknown range of values yielding a noticeable degradation of the prediction performance of the VM model based on the complete and MW DS.

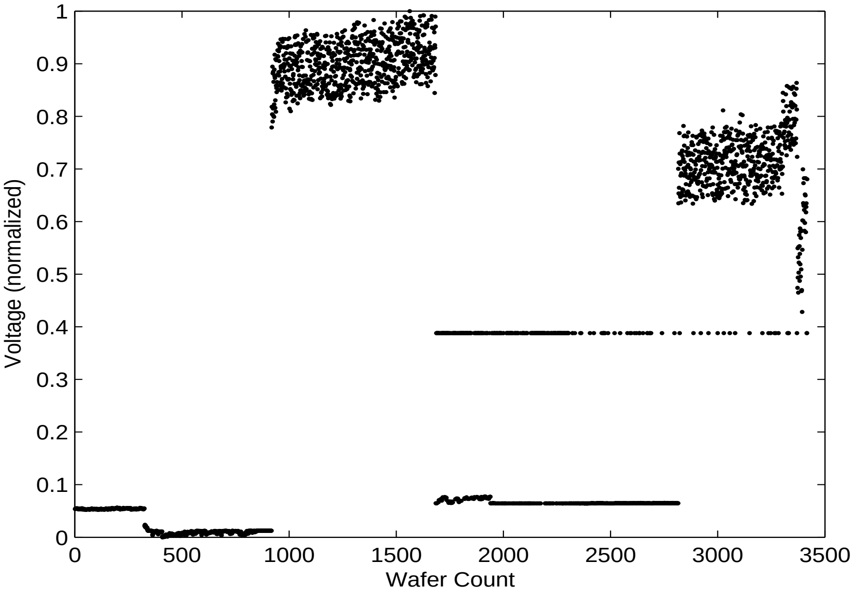


Figure 8.4: Due to readjustment of several process parameters for EQ_1 related to a periodical maintenance, Voltage_Chuck_1 as one of the crucial input features selected by ERBE FS shows a significant shift to a so far unknown range of values yielding a noticeable degradation of the prediction performance of the VM model based on the complete and MW DS.

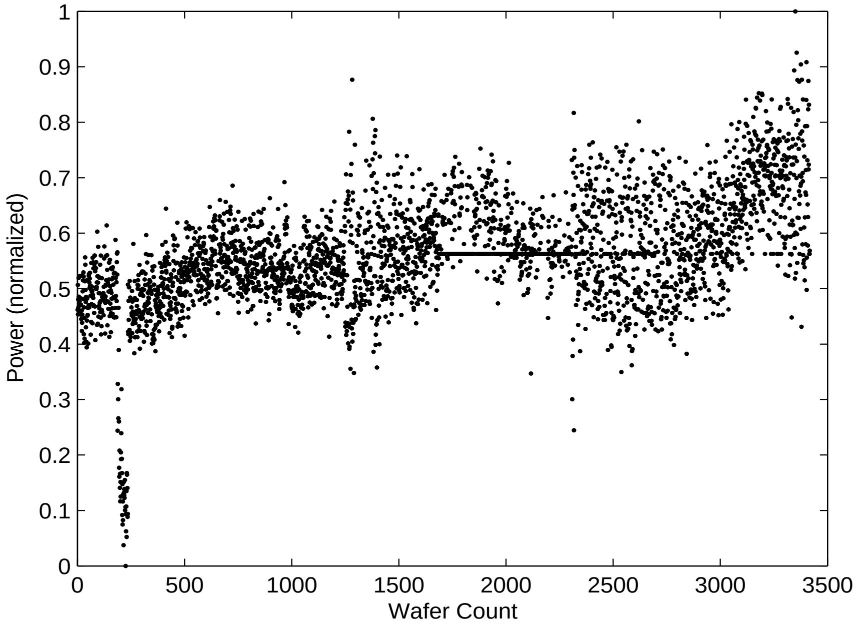


Figure 8.5: For comparison, Power_TS_1 of EQ_1 shows the typical variation as one of the crucial input features selected by ERBE FS for the two-and-a-half years DS not perceptibly affected by the maintenance intervention.

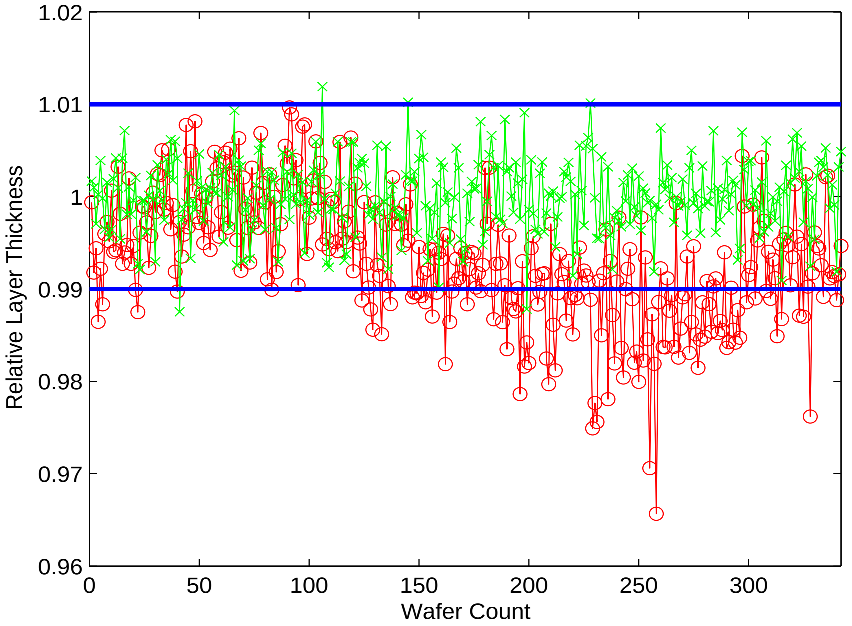


Figure 8.6: For EQ_1 the prediction (red) of the relative layer thickness for all wafers of the second half of the independent test DS for the VM model trained on the *updated MW* DS based on three months time period (i. e. first half of test DS) shows significant improvement compared to prediction in figures 8.1 and 8.2. However, minor degradation from ~ wafer 120 can be observed and is future subject to enhance the MW approach. The measured target (green) is provided with UCL and LCL (blue).

8.1.2 Equipment 2

Training DS	CV(RMSE)	CV(MAE)	R^2	Sen.	# Outliers	# Test	# Train
Complete	1.95	1.83	-18.32	0.17	6	415	1662
MW	3.90	3.84	-76.31	0	6	415	417
updated MW	0.57	0.47	-1.24	0.25	4	187	209

Table 8.2: Results for EQ_2 for complete, MW and updated MW DS: Evaluation of accuracy by CV(RMSE), CV(MAE) and R^2 as well as reliability by sensitivity including the number of outliers and instances within test and training DS.

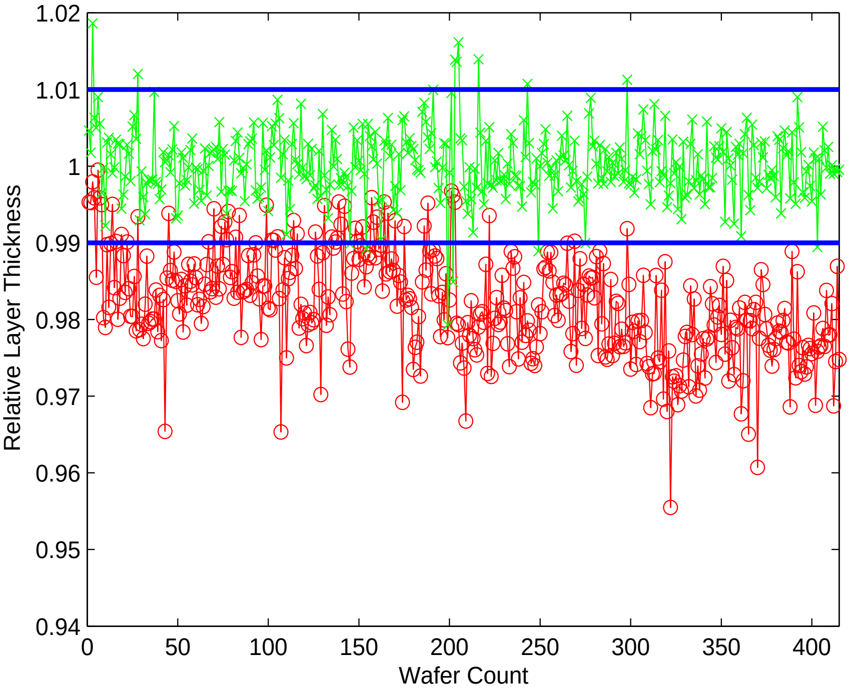


Figure 8.7: For EQ_2 the prediction (red) of the relative layer thickness for all wafers of the independent test DS for the VM model trained on the *complete* DS based on two-years time period shows significant degradation from ~ wafer 10. The measured target (green) is provided with UCL and LCL (blue).

Similar to EQ_1, the primary training of the VM prediction model for EQ_2 obviously failed for both used DSs (i. e. complete & MW) identified by all evaluation measures (i. e. RMSE, MAE, R^2 & sensitivity) and clearly visible in the figures 8.7 and 8.8.

Again, even though the obtained RMSE of 1.9 % and 3.9 % is smaller than for EQ_1 with superior results for the complete DS and comparable to some results of various publications stated in section 4.2, it is also not considerable for implementation of VM since control limits of process specifications for nanoscale designed high-tech products in SM constantly shrink. Feasibility and applicability of VM also depends on predictions with an accuracy of $< 1\,\%$. Right at the transition from the training to the test DS the equipment was as EQ_1 subject to a

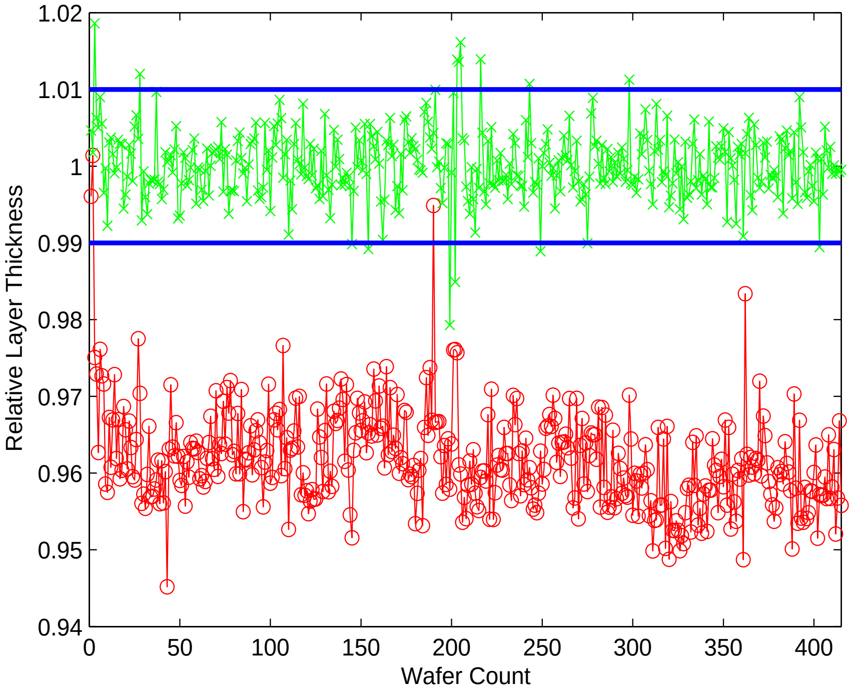


Figure 8.8: For EQ_2 the prediction (red) of the relative layer thickness for all wafers of the independent test DS for the VM model trained on the *MW* DS based on 6 months time period shows significant degradation. The measured target (green) is provided with UCL and LCL (blue).

major maintenance intervention as described in subsection 2.3.2. The required readjustment of the process parameter Voltage_Chuck_1 as one of the eight most important features selected by the ERBE algorithm (cf. table 8.7) can be observed for the last ~400 instances as displayed in figure 8.9 which illustrates the normalized voltage of the input feature Voltage_Chuck_1 for the combined training and test DS. Clearly visible is the change at wafer number ~ 1600 where the values are adjusted to almost normalized 0 after the periodical maintenance. The significant change of only one of the most important input features causes the VM model to perceptibly degrade and to fail to accurately and reliably predict the relative layer thickness. The adjustment of process parameters can impressively differ from one equipment to another (compare Voltage_Chuck_1 for EQ_1 & EQ_2).

The prediction conducted for the updated MW DS (cf. subsection 7.2.2) yields improved results illustrated in figure 8.10. As for EQ_1 a significant improvement with regard to all evaluation measures is achieved and corroborates the effectiveness of the MW approach after a periodical maintenance also motivating investigations regarding its further enhancement. Even though some outliers are not detected and some false outliers are predicted the very small RMSE of 0.57 % is a remarkable result for the prediction based on the updated MW DS.

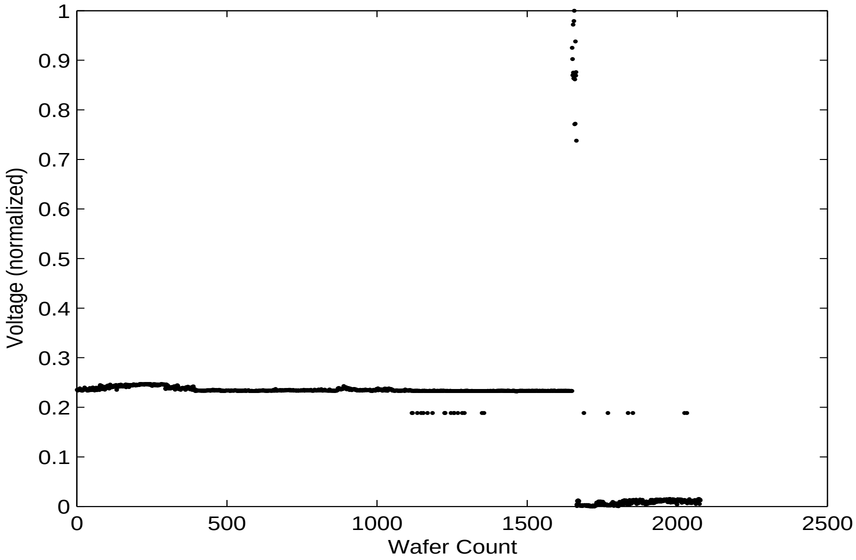


Figure 8.9: Due to readjustment of several process parameters for EQ_2 related to a periodical maintenance, Voltage_Chuck_1 as one of the crucial input features selected by ERBE FS shows a significant shift to a so far unknown range of values yielding a noticeable degradation of the prediction performance of the VM model based on the complete and MW DS.

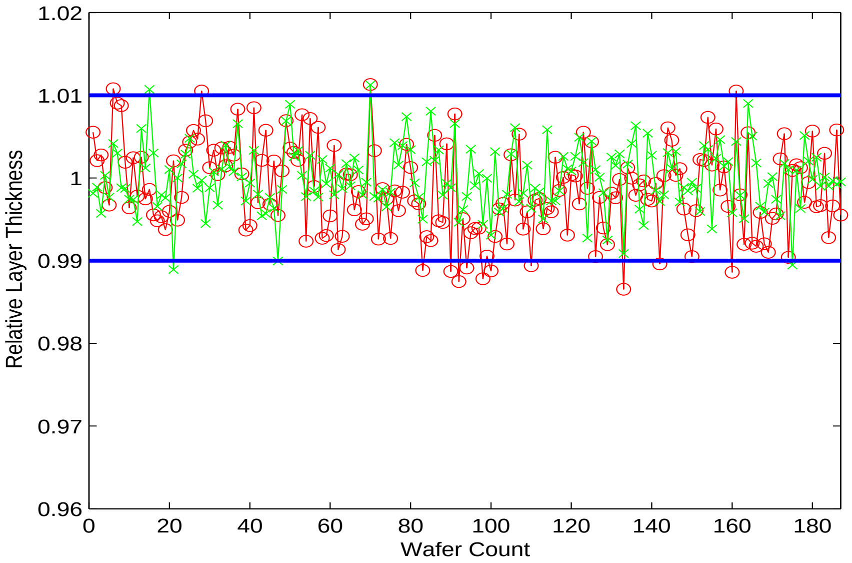


Figure 8.10: For EQ_2 the prediction (red) of the relative layer thickness for all wafers of the second half of the independent test DS for the VM model trained on the *updated MW* DS based on three months time period (i. e. first half of test DS) shows significant improvement compared to prediction illustrated in figures 8.7 and 8.8. In contrast to the previous updated MW evaluation (cf. figure 8.6) good prediction performance (i. e. RMSE = 0.57 %) is achieved for the all data approving the effectiveness of updated MW approach. The measured target (green) is provided with UCL and LCL (blue).

8.1.3 Equipment 3

Training DS	CV(RMSE)	CV(MAE)	R^2	Sen.	# Outliers	# Test	# Train
Complete	0.93	0.68	-5.35	0	4	468	1874
MW	0.83	0.66	-4.08	0.50	4	468	470

Table 8.3: Results for EQ_3 for complete and MW DS: Evaluation of accuracy by CV(RMSE), CV(MAE) and R^2 as well as reliability by sensitivity including the number of outliers and instances within test and training DS.

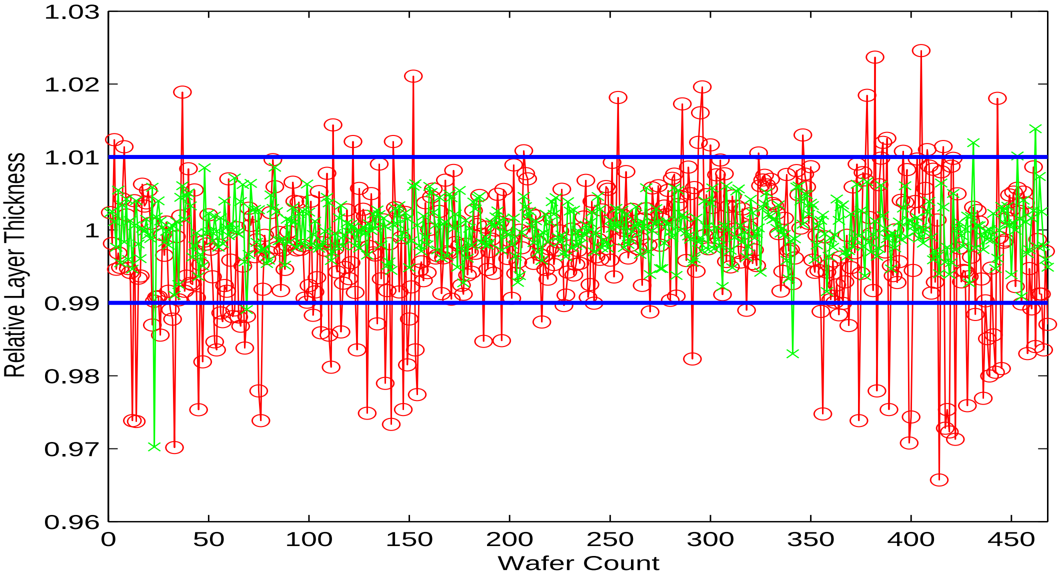


Figure 8.11: For EQ_3 the prediction (red) of the relative layer thickness for all wafers of the independent test DS for the VM model trained on the *complete* DS based on two-years time period shows mainly acceptable results but including too many predicted outliers. The measured target (green) is provided with UCL and LCL (blue).

EQ_3 shows no offset in any of the relevant features related to a periodical maintenance intervention close to the transition from the training to the test within the two-and-a-half years for the VM prediction model trained on the complete and MW DS. The results for the evaluation measures R^2, MAE and RMSE in table 8.3 are very promising but come along with the drawback of poor sensitivity visible in the figures 8.11 and 8.12. None of the four outliers are detected by the VM model for the complete DS but two (50 % sensitivity) are detected by the VM model based on the MW DS giving another hint to further focus on the MW approach. In addition, the prediction for the smaller MW DS yields slightly superior results in terms of RMSE. Despite these good results, both VM models need to be used carefully due to the significantly increased variance of the prediction compared to the observed target. The SVR-based VM model could further be investigated for suffering from possible overfitting since the predictions tend to strike above UCL or below LCL (blue). Nevertheless, falsely predicted outliers would lead only in combination with similar predictions from other VM models (e. g. NN, M5' – cf. section 5.2.2)

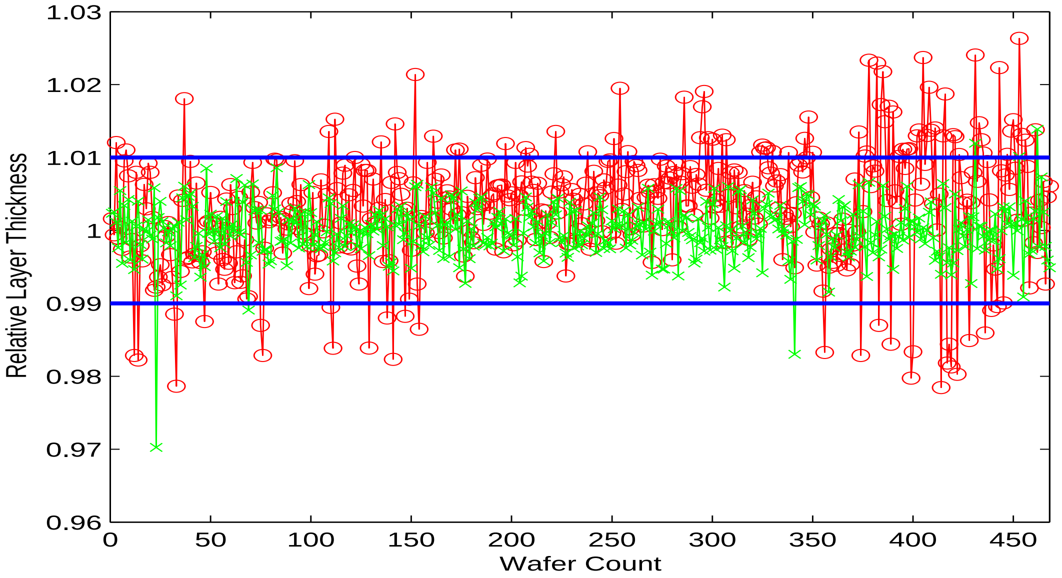


Figure 8.12: For EQ_3 the prediction (red) of the relative layer thickness for all wafers of the independent test DS for the VM model trained on the MW DS based on 6 months time period shows mainly acceptable results but including too many predicted outliers. The measured target (green) is provided with UCL and LCL (blue).

to repeated physical metrology and are still within a range of $< 3\,\%$ of the relative layer thickness. A slight expansion of UCL and LCL might be tolerable if the centered process capability index is not significantly decreased (cf. subsection 9.1.1).

8.1.4 Equipment 4

Training DS	CV(RMSE)	CV(MAE)	R^2	Sen.	# Outliers	# Test	# Train
Complete	0.67	0.53	-2.33	0.75	4	745	2980
MW	0.59	0.46	-1.59	0.75	4	745	746

Table 8.4: Results for EQ_4 for complete and MW DS: Evaluation of accuracy by CV(RMSE), CV(MAE) and R^2 as well as reliability by sensitivity including the number of outliers and instances within test and training DS.

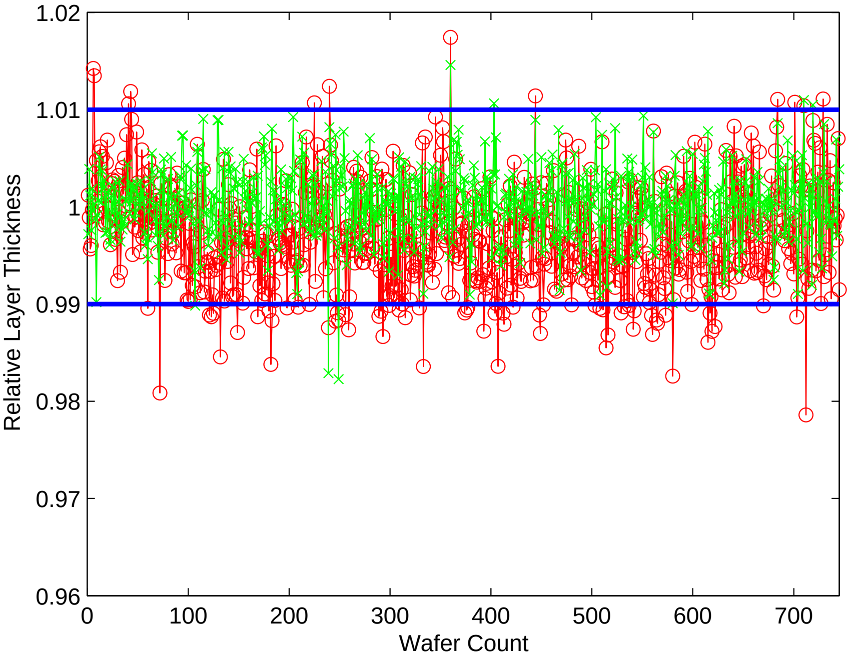


Figure 8.13: For EQ_4 the prediction (red) of the relative layer thickness for all wafers of the independent test DS for the VM model trained on the *complete* DS based on two-years time period shows good results with few significant deviations. The measured target (green) is provided with UCL and LCL (blue).

Similar to EQ_3, EQ_4 shows no offset in any of the relevant features related to a periodical maintenance intervention close to the transition from the training to the test within the two-and-a-half years for the VM prediction model trained on the complete and MW DS. The very good results for the evaluation measures R^2, MAE and RMSE $< 0.7\,\%$ as well as for reliability by means of sensitivity of detected outliers in table 8.4 are illustrated in the figures 8.13 and 8.14. Three out of four outliers (75 % sensitivity) are detected in the independent test DS by both VM models corroborating a demonstrative use case of VM in SM with the largest investigated DS including 3725 instances (i. e. wafers). Regarding R^2, MAE and RMSE, the prediction for the smaller MW DS yields slightly superior results whereat in addition the VM model trained on the complete DS produces some more small outliers (i. e. $< 2\,\%$ relative layer thickness) below the LCL (blue). Out of more than 700 tested instances the few falsely predicted outliers especially

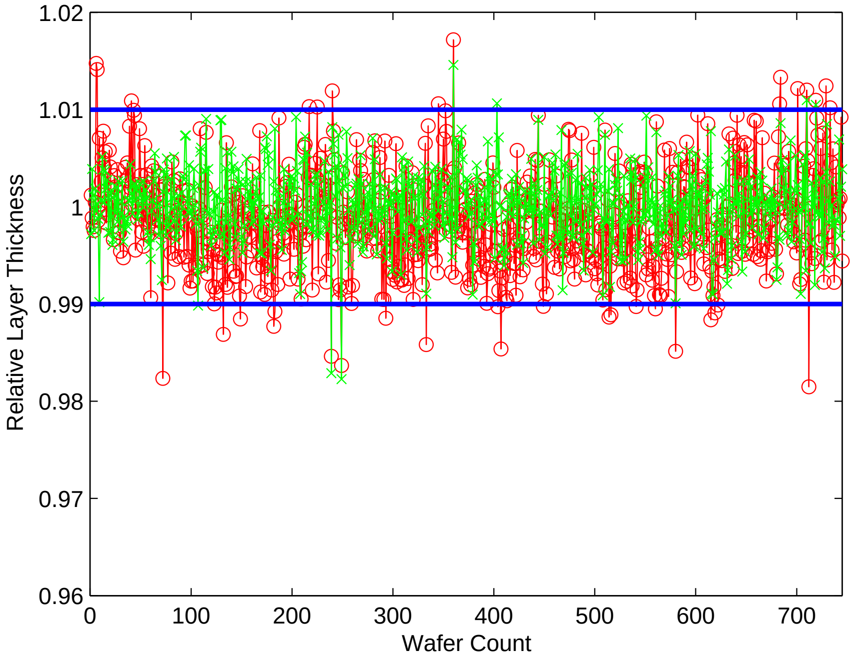


Figure 8.14: For EQ_4 the prediction (red) of the relative layer thickness for all wafers of the independent test DS for the VM model trained on the *MW* DS based on 6 months time period shows good results with few significant deviations. The measured target (green) is provided with UCL and LCL (blue).

in the approach based on the MW DS are overall acceptable since an almost optimal solution can be expected due to the fact that in statistical tests always a tradeoff between type one (false positive) and type two (false negative) errors needs to be found (cf. subsection 3.6.3). Hence, the constantly quite remarkable prediction performance for EQ_4 in figure 8.13 and especially in figure 8.14 demonstrates an applicable, targeted and desirable scenario for VM in high-mixture-low-volume SM.

8.1.5 Equipment 5

Training DS	CV(RMSE)	CV(MAE)	R^2	Sen.	# Outliers	# Test	# Train
Complete	0.58	0.46	-2.82	1	3	483	1934
MW	0.52	0.41	-2.11	0.67	3	483	485

Table 8.5: Results for EQ_5 for complete and MW DS: Evaluation of accuracy by CV(RMSE), CV(MAE) and R^2 as well as reliability by sensitivity including the number of outliers and instances within test and training DS.

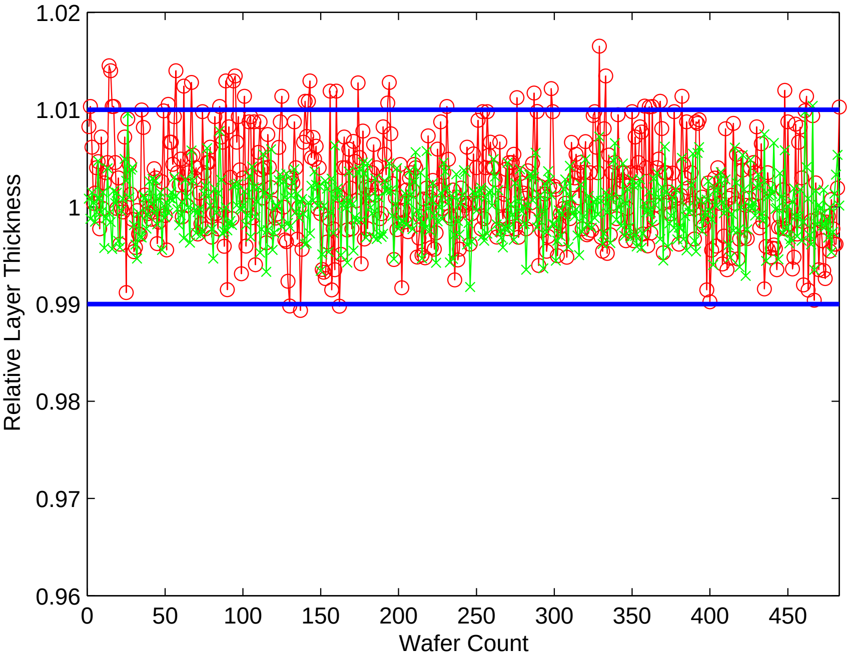


Figure 8.15: For EQ_5 the prediction (red) of the relative layer thickness for all wafers of the independent test DS for the VM model trained on the *complete* DS based on two-years time period shows good results with few deviations. The measured target (green) is provided with UCL and LCL (blue).

Also EQ_5 was not subject to a periodically maintenance intervention close to the transition from the training to the test within the two-and-a-half years time period for the VM prediction model trained on the complete and MW DS. The quite excellent results for the identified evaluation measures R^2, MAE and RMSE $< 0.6\,\%$ as well as for reliability by means of sensitivity of detected outliers in table 8.5 are visualized in the figures 8.15 and 8.16. At least two out of three outliers (67 % sensitivity) are detected in the independent test DS by both VM models again as for EQ_4 corroborating a demonstrative use case of VM in SM. With regard to R^2, MAE and RMSE, the prediction for the smaller MW DS yields slightly superior results whereat in addition the VM model trained on the complete DS produces some more small outliers (i. e. $< 2\,\%$ relative layer thickness) above the UCL (blue). Again, out of more than 450 tested instances the few falsely predicted outliers especially in the approach based on the MW DS are

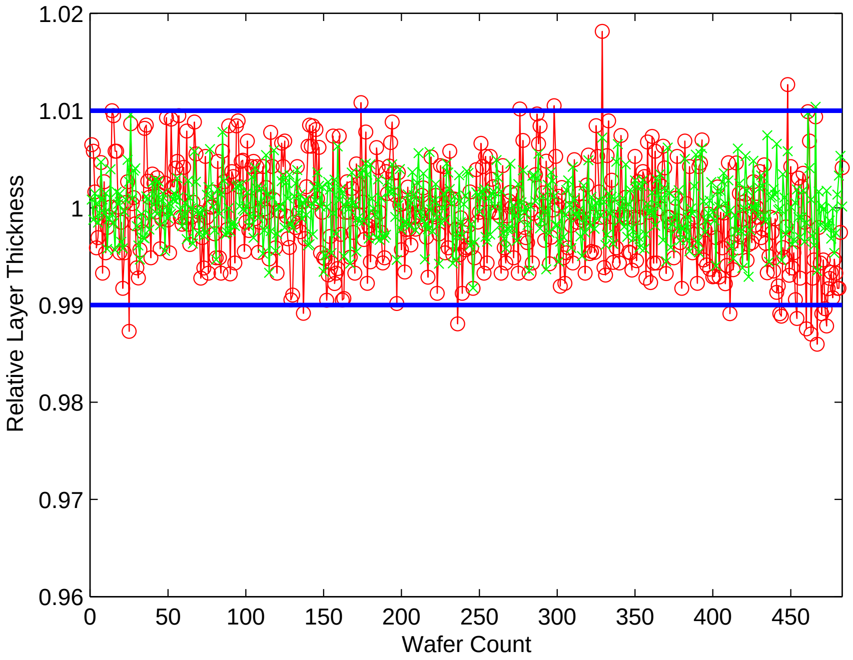


Figure 8.16: For EQ_5 the prediction (red) of the relative layer thickness for all wafers of the independent test DS for the VM model trained on the MW DS based on 6 months time period shows good results with few significant deviations. The measured target (green) is provided with UCL and LCL (blue).

totally acceptable and a solution close to the optimum can be expected. The investigation of VM for SM yields the overall best results for EQ_5 according to both, accuracy and reliability measured by RMSE of 0.6 % and sensitivity of 1, respectively. Once again, the constantly quite remarkable prediction performance for EQ_5 in figure 8.15 and especially in figure 8.16 demonstrates an applicable and desirable scenario as targeted for VM in high-mixture-low-volume SM.

8.2 Smart Feature Selection

Following the results for the experimental evaluation of the advanced VM system as first essential part of the present thesis, the newly developed ERBE FS algorithm is investigated as outlined in section 7.3. At first, the generic approach of ERBE FS is approved for different production equipment (i. e. AMAT Centura & AMAT Producer) running different manufacturing processes (i. e. HDP CVD & PECVD) which are detailed in section 2.4. These results demonstrate the potential of the generic ERBE FS for VM to tackle the essential problems of corporate-wide efficient deployment of VM in SM, scalability, knowledge discovery and highest possible accuracy stated in subsection 4.1.1. Subsequently, the ERBE FS technique is compared to the RELIEF filter and LOO wrapper FS methods (cf. section 4.3 & section 6.3).

8.2.1 ERBE Feature Selection

For the evaluation of the FS processing and the further investigation in terms of the remaining optimization potential, the ERBE FS algorithm (cf. section 6.5) is initially executed on available data of the AMAT Centura production equipment. Subsequently, the ERBE FS is also performed on productive data of the AMAT Producer equipment to confirm the generic approach of the developed algorithm.

ERBE Feature Selection for AMAT Centura

The following charts 8.17 – 8.25 illustrate the pareto of the approximately least important 20 % of features for each feature subset selected in the individual ERBE stages. The y-axis indicates how often a feature is selected whereas along the x-axis the corresponding feature names are listed (cf. appendix A.4). The least important 10 % features (red) are removed after each ERBE stage which are for the LOO stages (1–3 and 7–9) the ones selected the most as less significant features as average over ten cycles per stage and for the GA stages (4–6) the features most often omitted from the gene of the final best individual of as average over 25 cycles per stage to achieve more differentiation (cf. section 6.5). In case of several features left with the same amount of least important features, randomly some are removed (red) and others left in the feature subset (yellow). Artificial features (burnt orange) are not removed (cf. section 7.3.2) even in case of performing worse than other features. Features in blue belong to the least important 20 % but not the least important 10 % (except of artificial features) and are thus kept in the feature subset. For reference, in order to access the level of differentiation between the least important features the average and the one sigma intervals are provided as well.

ERBE Stage 1: The results in terms of the least important features for the first ERBE stage are displayed in figure 8.17. Already three out of the five artificial features are selected frequently as less significant whereat in fact it is interesting to recognize that their correlated complements (i. e. Artificial_2_ran for Artificial_4_cor & Artificial_5_cor for Artificial_3_ran) are not selected within the least important 20 % features. Thus, an original feature set containing many inferring features is most likely where even after executing LOO FS still many noisy variables are included. Features of various categories are revealed as dispensable while the range how often features are selected is widely spread from nine down to one. Furthermore, the high standard deviation of three around the average of four approves the good distinction within least important features. Finally, Temperature_Dome_5 is clearly recognized as most unimportant real feature.

ERBE Stage 2: In figure 8.18, the features for the second ERBE stage are less but still significantly spread as indicated by the standard deviation of two whereof two features selected two times are chosen randomly for elimination. Again, the same three artificial features are present as in the first ERBE stage and process parameter Flow_Oxygen_2 is obviously unimportant. Two remaining artificial features are not within the 15 % of the least important features and thus due to transition criterion 1.2 preventing few very frequently selected artificial features (cf.

subsection 6.5.1) the next ERBE stage is still executed with LOO FS.

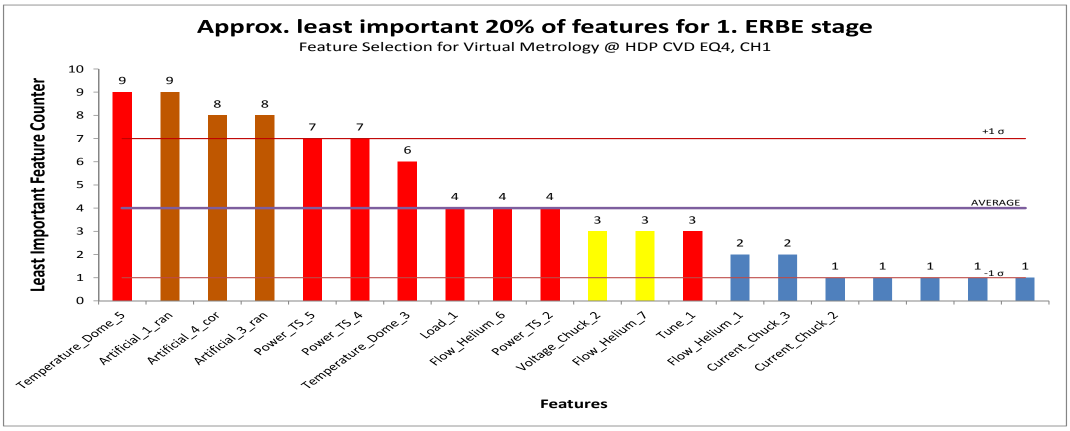


Figure 8.17: ERBE stage 1 indicating features selected within approximate least important 20 % and yielding a CV(RMSE) of 0.5919 is designed for fast elimination of features contributing mainly noise like artificial features (burnt orange) by LOO FS. Least important and removed real features (red) are distinguished very well from others randomly surviving features (yellow) or more important ones (blue).

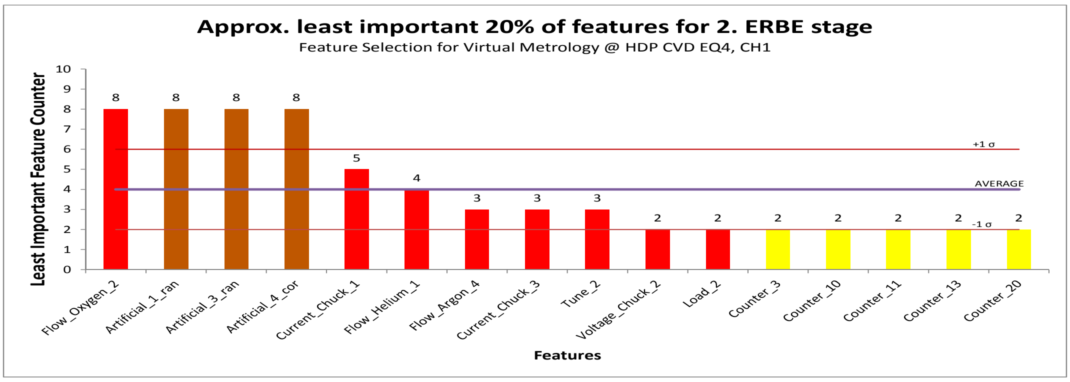


Figure 8.18: ERBE stage 2 indicating features selected within approximate least important 20 % and yielding a CV(RMSE) of 0.5836 is designed for fast elimination of features contributing mainly noise like artificial features (burnt orange) by LOO FS. Least important and removed real features (red) are distinguished quite well from others randomly surviving features (yellow).

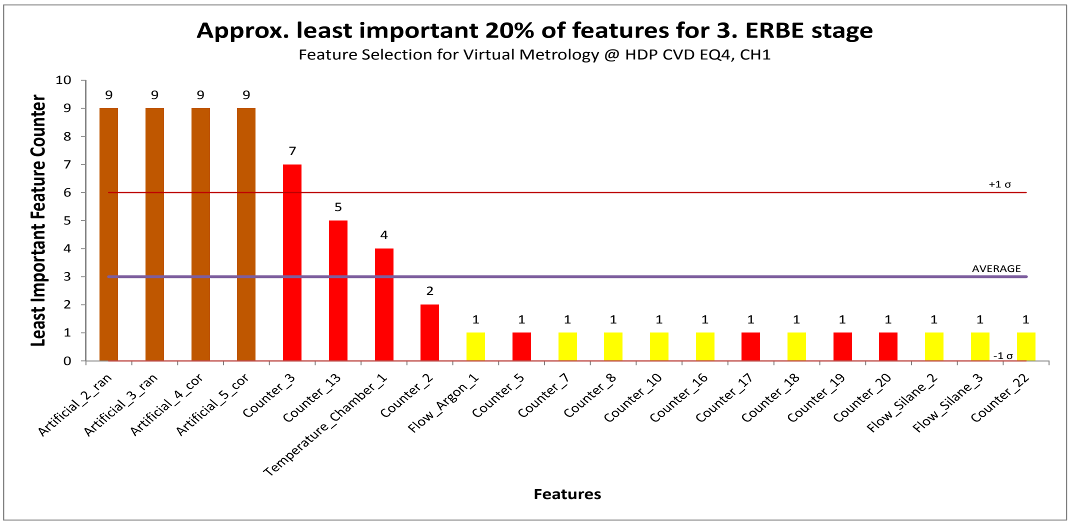


Figure 8.19: ERBE stage 3 indicating features selected within approximate least important 20 % and yielding a CV(RMSE) of 0.5789 is designed for fast elimination of features contributing mainly noise like artificial features (burnt orange) by LOO FS. Least important and removed real features (red) are distinguished mainly very well from others randomly surviving features (yellow).

ERBE Stage 3: The outcome of the third ERBE stage in figure 8.19 demonstrates for the first time the precise distinction by the algorithm between artificial features (i. e. Artificial_2_*–Artificial_5_*') and real features even though Artificial_1_ran is still not within 20 % of the least important features. The highly correlated artificial features pairs (i. e. Artificial_2_ran & Artificial_4_cor and Artificial_5_cor & Artificial_3_ran) are all selected exactly 9 times as least significant variable indicating a good result of LOO FS. Thus, the transition from ERBE part I performing fast feature elimination by LOO FS to ERBE part II optimizing the feature subsets by GA FS is observed after this stage (cf. subsection 6.5.2 algorithm 4). The number of least important real features quickly descends from 7 and half of the 8 eliminated features are selected only once and chosen again randomly for elimination. Many of the features which are selected only once belong to the category counter indicating the possible equivalent contribution of these features. As only the artificial features are selected very often and the remaining features rather rare, the average observed in the third ERBE stage dropped to 3 and the standard deviation increased again to 3. The worst feature (Counter_3) is one of the supervisory process parameters (e. g. counting processed wafers) to estimate the degradation of equipment hardware and the remaining time until the next maintenance intervention is needed. Transition criteria 1.1 (1.9 = 4 artificial features each 9 times selected / 5 real features 7, 5, 4, 2 and 1 time selected) and 1.2 (4 out of 5 features) of the ERBE algorithm (cf. subsection 6.5.2) are both met and thus ERBE part II with GA FS is subsequently executed.

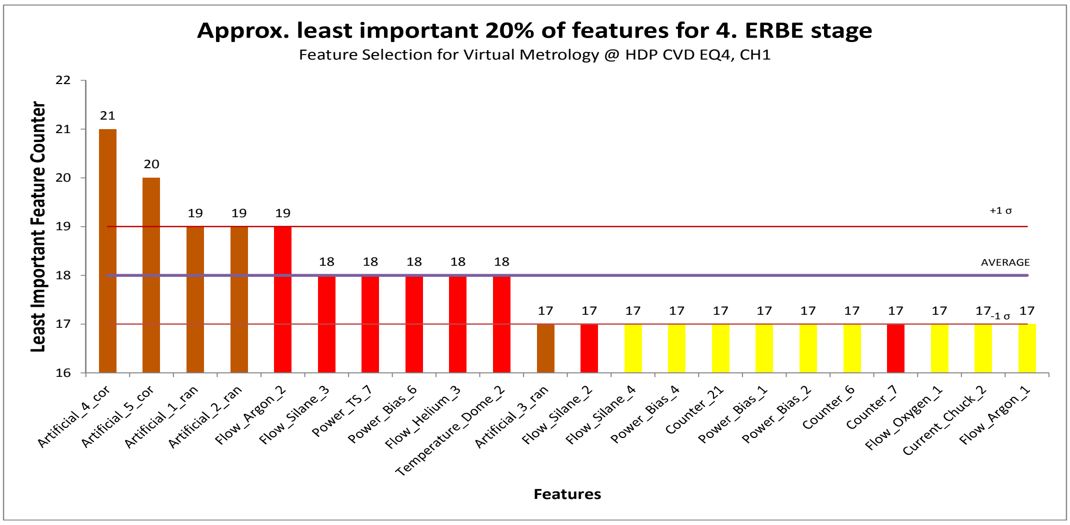


Figure 8.20: ERBE stage 4 indicating features selected within approximate least important 20 % and yielding a CV(RMSE) of 0.5452 is designed for feature subset optimization incorporating crucial interdependencies by GA FS. Least important and removed real features (red) are distinguished from others randomly surviving features (yellow) and at the first time all artificial features (burnt orange) are detected as unimportant.

ERBE Stage 4: In the first of the subsequent three ERBE GA stages of part II, all artificial features are present for the first time within the approximately least important 20 % features as indicated in figure 8.20, however Artificial_3_ran is still ranked below 6 real features. The average increased to 18 while the standard deviation decreased to one indicating a very small differentiation between the least important features. The fact of a small distinction between dispensable features is corroborated by quite frequent (i. e. 17 or more times out of 25 conducted cycles) exclusion of all least important features in the fourth ERBE stage from the final best individual optimized by GA FS. As before, features of various categories have been discovered as unimportant. Finally, transition criterion 2.2 (cf. subsection 6.5.1) is not met and therefore the next ERBE stage is executed again by GA FS.

ERBE Stage 5: Following the previous observation of overall frequently neglected features, figure 8.21 for ERBE stage 5 shows a y-axis range from 18 down to 5 where the most features are selected 5 to 8 times. The artificial features are clearly detected as useless input but again with interesting differentiation between only duplicated and thus perfectly correlated artificial features i. e. Artificial_2_ran vs. Artificial_4_cor and Artificial_3_ran vs. Artificial_5_cor yielding

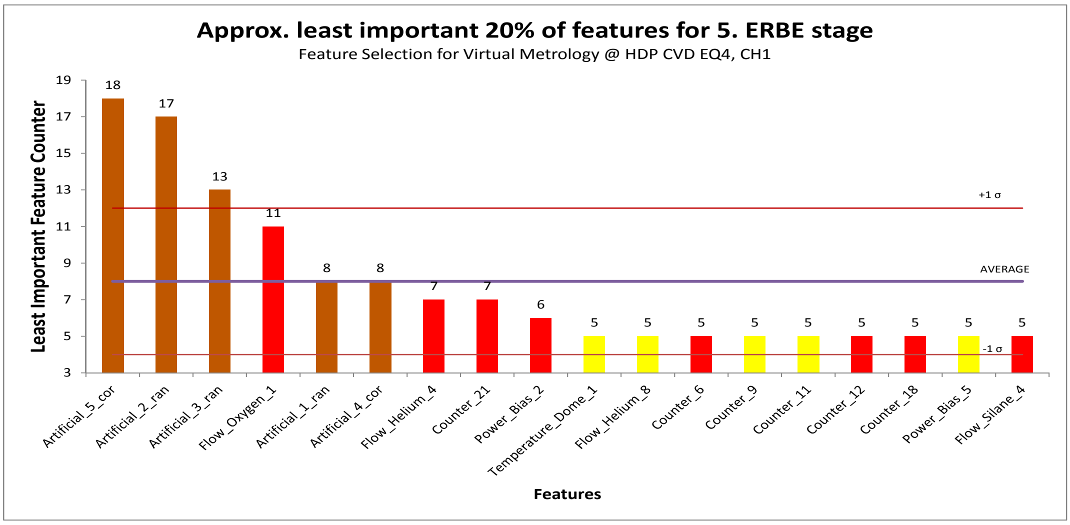


Figure 8.21: ERBE stage 5 indicating features selected within approximate least important 20 % and yielding a CV(RMSE) of 0.5716 is designed for feature subset optimization incorporating crucial interdependencies by GA FS. Least important and removed real features (red) are partially quite well distinguished from others randomly surviving features (yellow) whereas again all artificial features (burnt orange) are clearly detected as unimportant.

values of 17 vs. 8 and 13 vs. 18, respectively. The majority of the revealed least important features decreased the average from 18 down to 8 with an increased standard deviation of four due to the rarely chosen artificial features Artificial_2_ran and Artificial_5_cor. As before, features of every category are present with primarily eliminated features of the counter category and Flow_Oxygen_1 as noticeably least important feature. Finally, transition criterion 2.2 (cf. subsection 6.5.1) is not met and therefore the next ERBE stage is executed again by GA FS.

ERBE Stage 6: For the first time as figure 8.22 illustrates, all artificial features are clearly distinguished from the real features however again without matching the duplicated features Artificial_2_ran & Artificial_4_cor and Artificial_3_ran & Artificial_5_cor. An almost constant gradient and thus a linear line could be fitted quite well to describe the descending curve of counted least important features (especially the real ones) of the best individuals of the 25 cycles. Including artificial features, a slightly larger average (9) and smaller standard deviation (3) as in the previous ERBE stage emerged, but this metrics turns out to be very small (i. e. 8 & 1, respectively) if only the real features are considered. At last and similar to ERBE stage 3, counter process parameters (i. e. Counter_8 and Counter_16) are detected as very least important real features. Since both transition criteria (i. e. 2.1 & 2.2 - cf. subsection 6.5.1) are met the ERBE algorithm moves on to fine tuning feature optimization in part III. As the combination of

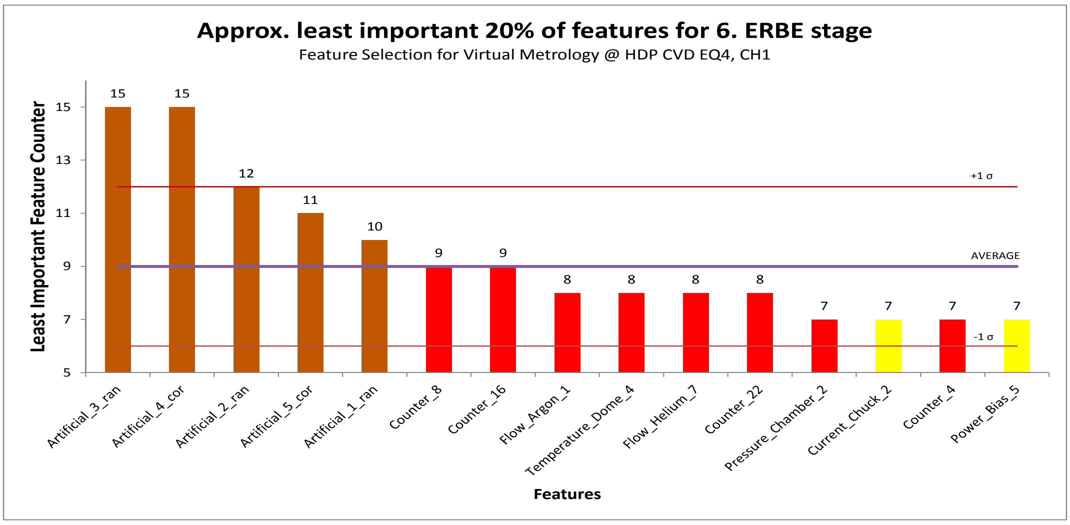


Figure 8.22: ERBE stage 6 indicating features selected within approximate least important 20 % and yielding a CV(RMSE) of 0.5841 is designed for feature subset optimization incorporating crucial interdependencies by GA FS. Least important and removed real features (red) are distinguished from others randomly surviving features (yellow) whereas again all artificial features (burnt orange) are clearly detected as unimportant.

clearly distinguished artificial features and similarly often selected real features is the first time noticeably observed for GA FS, the sixth ERBE stage indicates the transition at which feature subset optimization and global subspace search via heuristic GA FS moves to fine tuning feature optimization in part III (cf. subsection 6.5.2). In order to further improve the remaining feature subset while still ensuring high prediction performance, LOO FS is subsequently conducted again (cf. section 6.5).

ERBE Stage 7: With ERBE stage 7, part III of the ERBE algorithm starts to fine-tune and optimize the remaining feature subset via LOO FS with intended and initially achieved stronger feature differentiation as well as less computational effort for 10 LOO FS cycles compared to 25 GA FS cycles. The least important features with three or more counts are selected in addition to the five artificial features which again achieved the highest counts in the range of eight to nine counts. Feature Power_Bias_5 is eight times selected to be most dispensable while five out of the eight removed real features are selected three times explaining a comparably high standard deviation of three around an average of six. Thus, the targeted stronger feature distinction by LOO FS in figure 8.23 is corroborated by a considerable difference between irrelevant and less irrelevant features whereas no feature category is predominantly affected by elimination.

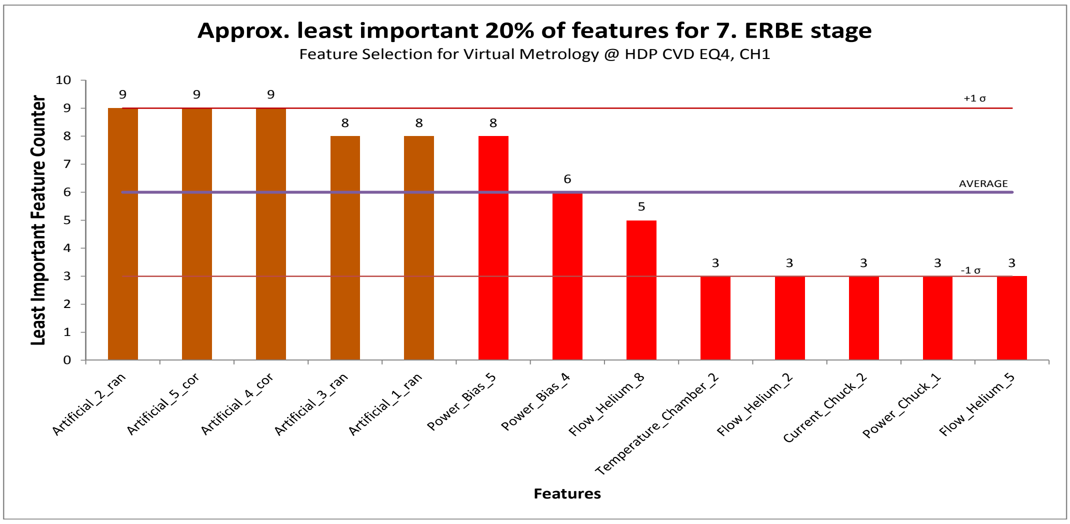


Figure 8.23: ERBE stage 7 indicating features selected within approximate least important 20 % and yielding a CV(RMSE) of 0.5649 is designed for fine tuning feature optimization incorporating crucial interdependencies by LOO FS. Least important and removed real features (red) are distinguished quite well whereas again all artificial features (burnt orange) are clearly detected as most unimportant.

ERBE Stage 8: According to figure 8.24 the eighth ERBE stage yields the same average and standard deviation (rounded) as the former stage whereas the range of the number of least important features increased as well as again the differentiation between real features which is explicitly intended in part III. It is also observable that the number of obtained least important features of duplicated artificial features Artificial_3_ran and Artificial_5_cor perfectly match in contrast to Artificial_2_ran and Artificial_4_cor. Interestingly, the artificial features are not grouped together as worst features anymore which might be due to overvaluing of Artificial_2_ran. This observation will be subject to further investigation since only 24 variables (including artificial ones) are present in the feature subset and a perfect correlation exists between these artificial features. Moreover, mostly (i. e. 5 out of 8) features of the category counter with Counter_9 as worst feature are eliminated compared to the previous stage where the categories are evenly distributed. Once more, a quite constant decrease of the amount of features in the descending curve is visible. The first time a small degradation yields a prediction accuracy above 0.6 % indicating loss of information and a tradeoff between model complexity and error (cf. section 3.3.4).

ERBE Stage 9: Figure 8.25 illustrates the ninth and last ERBE stage where only eleven real features are left as input and eight out of them are finally eliminated from the feature subset thus pushing the new ERBE FS with $rR = 10\,\%$ to the limit. Hence, the remaining three real

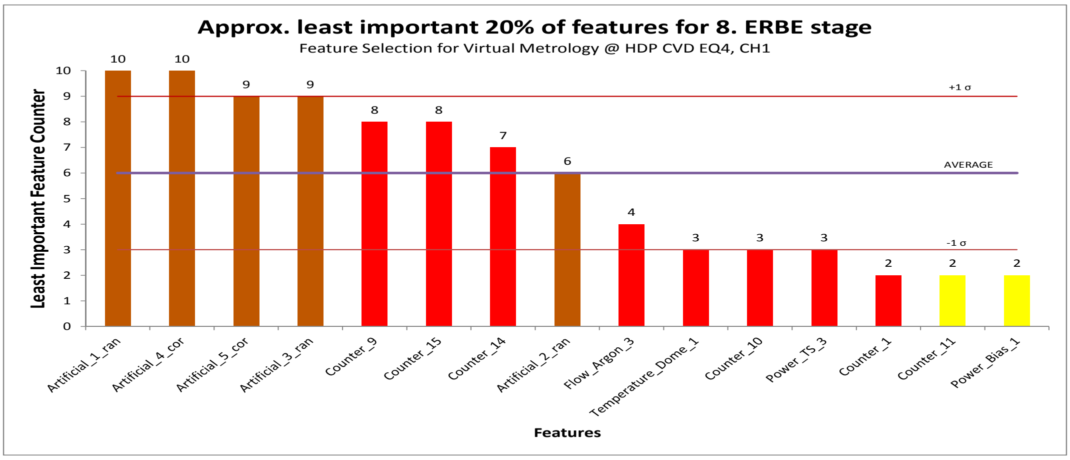


Figure 8.24: ERBE stage 8 indicating features selected within approximate least important 20 % and yielding a CV(RMSE) of 0.611 is designed for fine tuning feature optimization incorporating crucial interdependencies by LOO FS. Least important and removed real features (red) are distinguished very well from others randomly surviving features (yellow) whereas again the artificial features (burnt orange) are detected as unimportant.

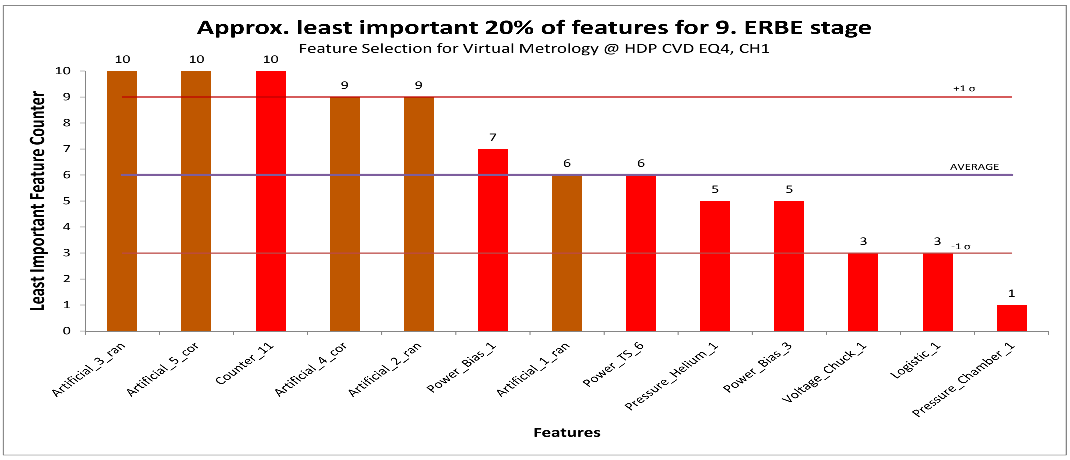


Figure 8.25: ERBE stage 9 indicating features selected within approximate least important 20 % and yielding a CV(RMSE) of 0.6234 is designed for fine tuning feature optimization incorporating crucial interdependencies by LOO FS. Least important and removed real features (red) are distinguished very well even though again not all artificial features (burnt orange) are clearly detected as most unimportant.

features not listed as least important ones are most important to achieve highest prediction performance. The artificial features are again not grouped together but the two artificial feature groups Artificial_3_ran & Artificial_5_cor and Artificial_2_ran & Artificial_4_cor perfectly match with nine and ten counts in each group. Counter_11 clearly turned out as dispensable feature with also ten counts. Again, the descending curve appears to be very straight keeping the average and standard deviation constant at the same values 6 and 3, respectively, as in the two previous stages. Thus, the intended and very well demonstrated differentiation of fine tuning feature optimization of part III (cf. algorithm 4) is corroborated once more while the steady small degradation again yields a prediction accuracy above 0.6 % indicating loss of information and a tradeoff between model complexity and error (cf. section 3.3.4).

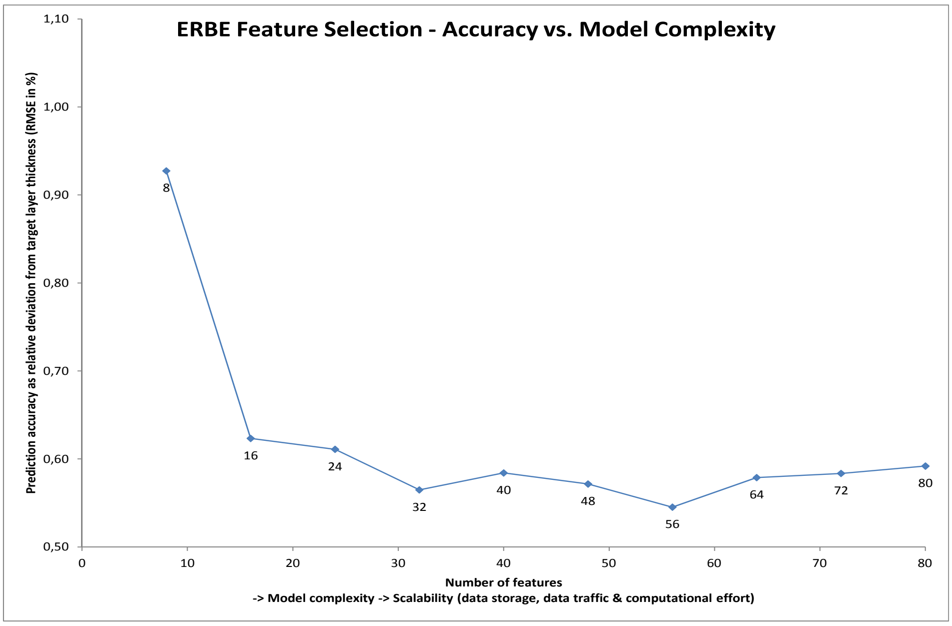


Figure 8.26: ERBE FS for AMAT Centura: Prediction performance for all ERBE stages and the final evaluation in terms of accuracy measured by the CV(RMSE) plotted against the number of features representing the challenge of scalability by means of the model complexity affecting data storage, data traffic and computational effort of the VM implementation.

In figure 8.26 the results of all ERBE stages for the AMAT Centura equipment (EQ_4, CH_1) are plotted against the number of labeled features representing the challenge of scalability (cf. subsection 4.1.1) by means of the model complexity (cf. Bias-Variance-Tradeoff section 3.3.4). In order to obtain a meaningful result after conducting the new ERBE FS algorithm the final feature subset consisting of the five artificial features and Pressure_Helium_2, Power_TS_1 as well as Flow_Silane_1 is also evaluated ten times with grid search optimization and yields a prediction performance of 0.9273 %. The significantly degraded prediction performance of > 0.3 % compared to the other ERBE stages with CV(RMSE) between 0.5452 up to 0.6234 impressively demonstrates the final loss of crucial information to accurately predict the target with only three real features left. Hence, the transition from very good to noticeably worse prediction accuracy is expected after the ninth (i. e. last) ERBE stage. The prediction accuracy as relative deviation from the target (i. e. LT) is measured by the CV(RMSE) to emphasize the desired recognition of outliers. Over all ERBE stages, an average accuracy of 0.5839 % is achieved with a maximum of 0.6234 % at stage nine and a minimum of 0.5452 % at stage four. The difference between these stages of only 0.0782 % indicates statistical variation related to different compositions of the training and validation DS due to shuffling of the instances as well as ever changing randomly created artificial features.

Investigation of Optimization Potential

The accuracy for all ERBE stages may improve if the artificial features are removed and only meaningful real features are included in the final feature subset. Hence, before the results of the new ERBE FS algorithm are compared to the RELIEF filter and LOO FS wrapper methods (cf. subsection 8.2.2), a manual feature subset optimization is performed around the turning point with 3 - 11 features left (i. e. $16 - 5 = 11$ excluding the artificial ones) at which the prediction performance significantly degrades to investigate the maximum possible accuracy for smallest feature subset.

At first, two HDP CVD process experts are interviewed and asked to provide a feature subset (i. e. Expert Selection – ES) only including all the features with expected high relevance and thus crucial for implementation of a VM system. These subsets containing 20 and 22 features are illustrated in figure 8.27 in addition to the ERBE FS curve as labeled purple triangles yielding a degraded prediction performance. Before any attempt is made to manually optimize the feature subsets selected by the ERBE FS algorithm, the resulting feature subset of the eighth ERBE stage is applied omitting the artificial features (i. e. only including the 11 real features) to train VM models again for the average of ten cycles also performing grid search. These results are displayed by a burnt orange diamond connected and belonging to the blue diamond for ERBE FS stage 8. The description in the legend $ERBE\_FS\_S8\_11$ of figure 8.27 indicates ERBE stage eight by S8 and the number of 11 input features.

Features obtained either from ES according to appropriate recommendations of the process experts or from the last 30 % (i. e. resulting from ERBE stage 7) are manually combined into feature subsets in order to investigate the improvement potential of the final feature subset selected by the ERBE FS algorithm. These manually optimized feature subsets are named FS_OPT_* where the asterisk is substituted by two numbers, the first one defining the number of features within the feature subset and the second one the index if different feature combinations with the same number of features within a subset are used. All different feature subset combinations are displayed as green dots in figure 8.27 (not labelled individually for the sake of clarity) and the composition of these feature subsets is provided in table 8.6 (cf. appendix A.4.1 for the full feature list).

Few features are adjusted at the equipment and as far as possible kept constant with the

Feature	Number	ES_22	ES_20	FS_OPT_3_1	FS_OPT_3_2	FS_OPT_3_3	FS_OPT_5	FS_OPT_7	FS_OPT_8_1	FS_OPT_8_2	FS_OPT_11_1	FS_OPT_12_1	FS_OPT_12_2	FS_OPT_13	FS_OPT_14_1	FS_OPT_14_2	FS_OPT_14_3	FS_OPT_16	FS_ERBE_S8_11
Temperature_Chamber_1	5	1	1																
Temperature_Chamber_2	6	1	1																
Flow_Helium_1	7		1								1	1	1	1	1	1	1	1	
Flow_Helium_2	8	1	1			1	1	1		1	1	1	1	1	1	1	1	1	
Pressure_Helium_1	9								1										1
Pressure_Helium_2	10								1										1
Pressure_Chamber_1	11	1	1					1	1	1	1	1	1	1	1	1	1	1	1
Temperature_Dome_1	13	1	1												1			1	
Temperature_Dome_2	14	1	1																
Temperature_Dome_3	15	1	1							1						1		1	
Temperature_Dome_4	16	1															1	1	
Temperature_Dome_5	17	1																	
Current_Chuck_1	18		1																
Power_Chuck_1	21		1																
Voltage_Chuck_1	22		1					1	1	1	1	1	1	1	1	1	1	1	1
Flow_Helium_3	24		1								1	1	1	1	1	1	1	1	
Flow_Helium_5	30	1	1								1	1	1	1	1	1	1	1	
Flow_Helium_6	31	1																	
Flow_Helium_7	32	1	1																
Counter_11	44																		1
Power_Bias_1	55		1									1		1	1	1	1	1	1
Power_Bias_3	57	1	1		1		1	1	1	1			1	1	1	1	1	1	1
Power_Bias_4	58	1																	
Power_Bias_5	59	1	1																
Power_TS_1	60			1			1	1	1	1	1	1	1	1	1	1	1	1	1
Power_TS_3	62	1									1	1	1	1	1	1	1	1	
Power_TS_4	63	1																	
Power_TS_5	64	1																	
Power_TS_6	65										1	1	1	1	1	1	1	1	1
Power_Bias_6	67	1																	
Flow_Silane_1	68		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Flow_Silane_3	70		1																
Logistic_1	73	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tune_1	74	1																	
Tune_2	75	1																	
Sum of features		22	20	3	3	3	5	7	8	8	11	12	12	13	14	14	14	16	11

Table 8.6: Feature subset composition to investigate the optimization potential of ERBE FS. Columns 1 & 2 state the features with their numbers (cf. full feature list in appendix A.4.1) and columns 3 & 4 (ES_22 & ES_20) present the features included by process experts in the performed ES. All columns beginning with 'FS_OPT_' (i. e. 5 to 19) provide the manually optimized feature subsets. The last column provide the feature subset composition after stage 8 of the ERBE FS algorithm without artificial features resulting in 11 input variables. The total sum of features per subset is added at the bottom.

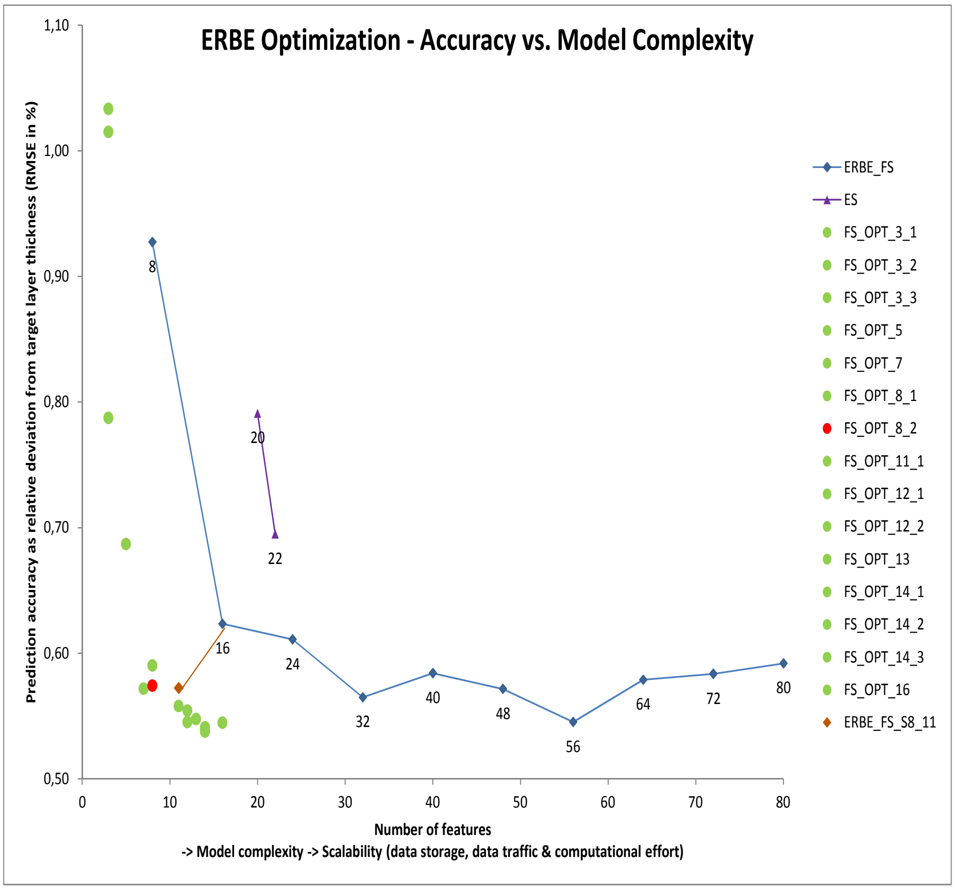


Figure 8.27: Investigation of Optimization Potential: The feature subset from ERBE stage eight ($ERBE\_FS\_S8\_11$) is evaluated and manually modified based on process expert recommendation (green & red dots) to obtain an optimized result in terms of accuracy and model complexity. For comparison, the result for the expert selected feature subsets is given (ES_20 & ES_22).

dome temperature as one representative. Although, a constant feature is neglected by any FS algorithm the strong recommendations of process experts to include the dome temperature (especially Temperature_Dome_3) into the investigation of the ERBE FS optimization potential is accepted since future changes of this feature may provide further valuable information.

The further recommendation of experienced process experts to substitute the two features Pressure_Helium_1 and Pressure_Helium_2 by Flow_Helium_2 and to incorporate the feature Temperature_Dome_3 are changes made to obtain $FS\_OPT\_8\_2$ (highlighted as red dot in figure 8.27) from $ERBE\_FS\_S8\_11$ (burnt orange) whereas feature subset FS_OPT_7 as green dot located left to FS_OPT_8_2 only lacks the dome temperature compared to the latter feature subset. Since the accuracy of the three feature subsets is almost the same and the model

Index	Feature	Category	Unit	Description
8	Flow_Helium_2	Gas flow	sccm	Helium gas flow into process chamber
11	Pressure_Chamber_1	Pressure	mTorr	Pressure within process chamber
15	Temperature_Dome_3	Temperature	°C	Temperature at ceramic dome
22	Voltage_Chuck_1	Voltage	V	Voltage applied to electrostatic chuck
57	Power_Bias_3	Power	W	DC-bias power applied by RF coil generator
60	Power_TS_1	Power	W	Top/Side power applied by RF coil generator
68	Flow_Silane_1	Gas flow	sccm	Silane gas flow into process chamber
73	Logistic_1	Logistics	-	Logistical Parameter

Table 8.7: Optimized ERBE FS feature subset to evaluate the VM system.

complexity defined by the feature subset size of quite the same order, it strongly corroborates the hypothesis of an almost optimal result revealed by the ERBE FS algorithm in terms of maximal accuracy for a minimal number of features. In order to consider the expert recommendations, *FS_OPT_8_2* is finally chosen as input feature subset to validate the advanced VM system (cf. chapter 5 and results in section 8.1) with its feature composition listed in table 8.7. The dimensionality reduction of more than 85 % of the input feature set containing initially 75 features (omitting the artificial ones) to finally 11 features in the *ERBE_FS_S8_11* feature subset also approves the remarkable performance of the ERBE FS algorithm.

ERBE Feature Selection for AMAT Producer

Regarding ERBE FS for the PECVD process on AMAT Producer (cf. section 2.4), only the curve to assess accuracy versus model complexity is illustrated below. The results of each ERBE stage are individually outlined and discussed (cf. appendix A.6).

Figure 8.28 plots the prediction performance in terms of relative deviation from the targeted layer thickness measured by CV(RMSE) against the number of features representing the model complexity and thus addressing the challenge of scalability (cf. section 4.1.1). In contrast to FS for the HDP CVD process on AMAT Centura a much higher number of features (i. e. 203 features including the five artificial ones) is initially available also determining with the reduction rate rR of 10 % an increase to 20/21 deleted features in each ERBE stage. ERBE part I performs LOO FS until a distinction between artificial and real features is achieved and the transition criteria (cf. subsection 6.5.1) are met which happened when 142 features are left after the third ERBE stage (cf. figure A.3). These first three steps clearly show a significant higher gradient and faster improvement of CV(RMSE) than the other ERBE stages of almost 0.5 % from ~ 1.8 % down to ~ 1.3 %. ERBE part II conducting feature subset optimization via GA FS still constantly improves the accuracy but in sum only by ~ 0.05 % until only 80 features are left when the transition point to switch to feature fine tuning again via LOO FS in ERBE part III is reached after three more steps (cf. figure A.6) similar to ERBE FS for the HDP CVD process on AMAT Centura. Only the last of the three ERBE stages in part III achieved a noticeable improvement of the accuracy (i. e. ~ 0.2 %) resulting in a final deviation of the targeted layer thickness of ~ 1.1 % with 20 features left in the final feature subset. Compared to ERBE FS for the HDP CVD process on AMAT Centura in figure 8.26 a significant and constant improvement

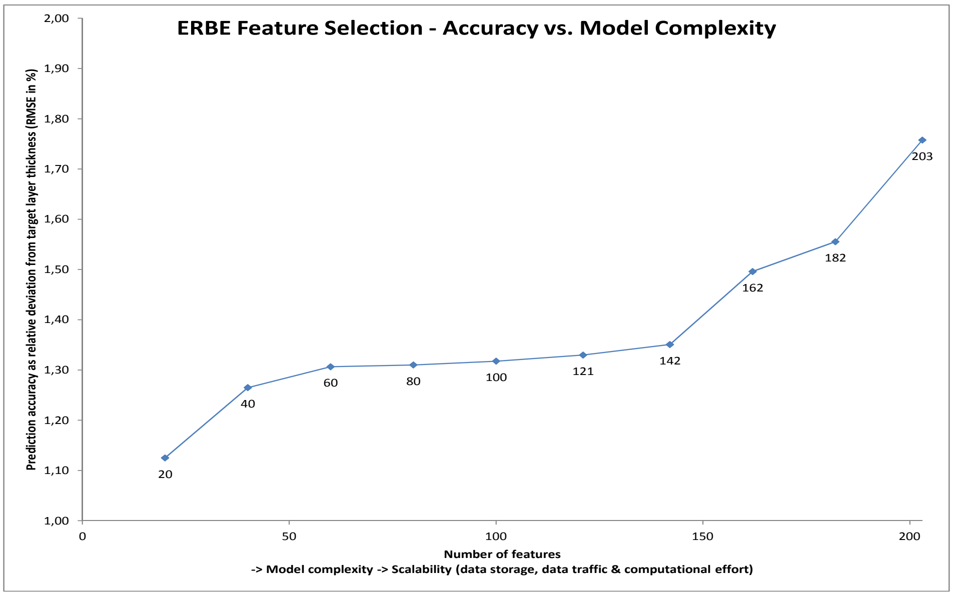


Figure 8.28: ERBE FS for AMAT Producer: Prediction performance for the initial feature set and all nine ERBE stages in terms of accuracy measured by the RMSE plotted against the number of features i. e. the model complexity affecting scalability by means of data storage, data traffic and computational effort of the VM implementation.

and strong monotone increasing prediction performance is observed whereat no degradation is visible at the end. Thus, noisy and redundant information is at first quickly and ongoing constantly removed from the feature set and it is reasonable to expect that further fine tuning of the final feature subset will on the one hand further increase the prediction performance and on the other hand will also end up with increasing loss of important information and therefore a reduction of accuracy again.

Hence, the optimum for the model complexity as described in section 3.3.4 and displayed in figure 3.2 is also located in a range of less than 15 features due to the fact that the final feature subset of 20 features still includes the 5 artificial ones. At last, the final feature subset comprising the 15 most important features is given in table 8.8 demonstrating a versatile mixture of features from all categories.

8.2.2 Comparison of ERBE, RELIEF and single LOO Feature Selection

A comparison of the ERBE FS algorithm with the well-established RELIEF FS filter and the LOO FS wrapper method is drawn where feature subsets comprising 3, 5, 8, 10, 12 and 15 features are investigated for the same DS as used for the evaluation of the ERBE FS. RELIEF (cf. section 4.3) as well as LOO FS (cf. section 6.3) are both feature ranking methods providing a feature list in decreasing order with the most important feature at the top. As outlined

Index	Feature	Category
4	Flow_Argon_4	Gas flow
11	Pressure_Chamber_4	Pressure
23	Power_Bias_2	Power
30	Logistic_1	Logistics
36	Power_Heater_5	Power
40	Power_Heater_9	Power
55	Temperature_Heater_15	Temperature
56	Temperature_Heater_16	Temperature
59	Temperature_Heater_19	Temperature
66	Flow_Helium_6	Gas flow
180	Pressure_Throttle_6	Gas flow
183	Time_2	Time
185	Time_4	Time
194	Time_13	Time
195	Time_14	Time

Table 8.8: Composition of final ERBE FS feature subset for PECVD on AMAT Producer.

in subsection 7.3.3, ten SVR models are trained for all three compared methods to relativize variation of randomly created artificial features and the final feature ranking is based on the average feature importance of these ten runs. Table 8.9 lists the first 15 ordered features for all three techniques together with the achieved CV(RMSE) for each feature subset with respect to their size in bold font.

No.	RELIEF		LOO FS		ERBE FS	
	Feature	CV(RMSE)	Feature	CV(RMSE)	Feature	CV(RMSE)
1	Power_TS_3		Power_TS_1		Pressure_Helium_2	
2	Power_TS_1		Pressure_Helium_2		Flow_Silane_1	
3	**Power_Bias_3**	**1.0788**	**Logistic_1**	**1.5124**	**Power_TS_1**	**0.9317**
4	Temperature_Dome_4		Pressure_Helium_1		Pressure_Chamber_1	
5	**Pressure_Chamber_1**	**0.7696**	**Power_TS_3**	**0.7915**	**Logistic_1**	**0.7403**
6	Temperature_Dome_1		Pressure_Chamber_1		Voltage_Chuck_1	
7	Power_Bias_1		Voltage_Chuck_1		Pressure_Helium_1	
8	**Flow_Helium_3**	**0.7208**	**Flow_Argon_3**	**0.7290**	**Power_Bias_3**	**0.6445**
9	Logistic_1		Power_Bias_3		Power_TS_6	
10	**Temperature_Dome_5**	**0.6553**	**Temperature_Dome_4**	**0.7080**	**Power_Bias_1**	**0.6330**
11	Power_Bias_6		Power_TS_6		Counter_11	
12	**Temperature_Dome_2**	**0.7226**	**Flow_Silane_1**	**0.6396**	**Counter_1**	**0.6287**
13	Flow_Helium_4		Power_Bias_1		Temperature_Dome_1	
14	Flow_Helium_1		Flow_Helium_5		Counter_10	
15	**Flow_Helium_8**	**0.7637**	**Power_Chuck_1**	**0.6366**	**Power_TS_3**	**0.6277**

Table 8.9: The first 15 ranked features of the RELIEF, LOO and ERBE FS techniques are listed together with the achieved CV(RMSE) for each corresponding evaluation of feature subsets comprising 3, 5, 8, 10, 12, and 15 features. The first column specifies the number of variables in each feature subset. The tested feature subsets indicated by their size are successively highlighted in bold font.

The results of the new ERBE FS algorithm (labeled in bold font by feature subset size) constantly outperform the strong monotone improving LOO FS wrapper and the RELIEF filter approach degrading for feature subset 12 and 15. Figure 8.29 clearly illustrates the superior

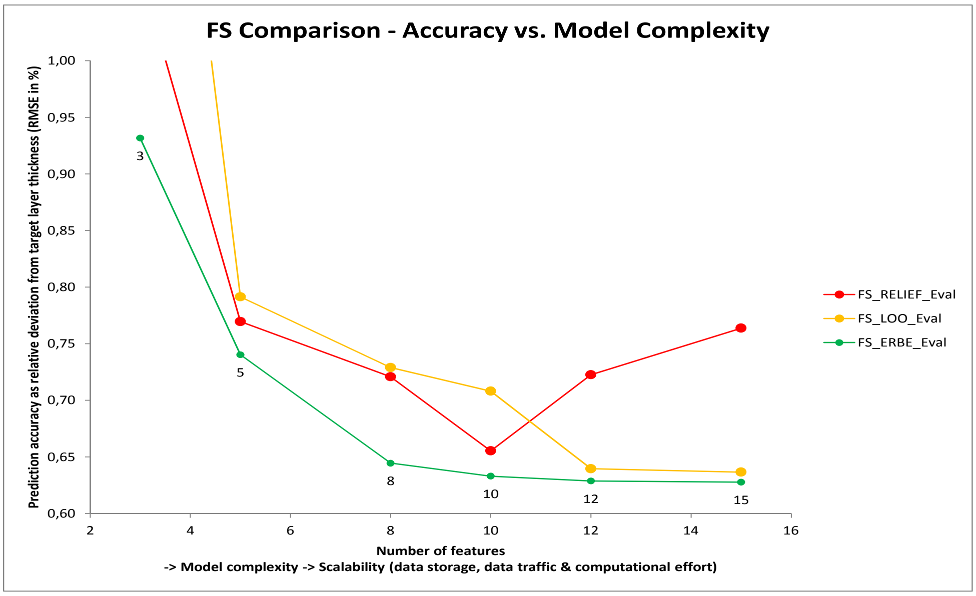


Figure 8.29: Comparison of RELIEF (red), LOO (orange) and ERBE FS (green) techniques by means of prediction accuracy as relative deviation form target layer thickness (RMSE in %) for feature subsets sizes 3, 5, 8, 10, 12 and 15. The results for the new ERBE FS algorithm (labeled by feature subset size) constantly outperform the strong monotone improving LOO FS wrapper and the RELIEF filter degrading for feature subset 12 and 15. The poor results for the feature subsets with three input variables for RELIEF FS (i. e. CV(RMSE) = 1.079) and LOO FS (i. e. CV(RMSE) = 1.5124) are not displayed to maintain a meaningful figure by appropriate scaling.

prediction performance for ERBE FS which already achieves with three input variables a prediction accuracy in terms of deviation from the target layer thickness $< 1\,\%$. The poor results for the feature subsets with three input features for RELIEF FS (i. e. CV(RMSE) = 1.079) and LOO FS (i. e. CV(RMSE) = 1.5124) are not displayed to maintain a meaningful figure by appropriate scaling. It is depicted very well that around the number of eight input features the curve of ERBE FS flattens and the prediction performance in terms of accuracy only improves marginally. While the RELIEF filter yields good results up to the input of ten features it significantly degrades for more features thus losing reliability due to the fact that the other methods still improve the prediction performance. The LOO FS wrapper constantly decreases the CV(RMSE) and finally achieves results close to the ones of ERBE FS for 12 and 15 input features. Nevertheless, a noticeable gap exists between both curves with the ERBE FS achieving superior results until these feature subset sizes are reached. As it is revealed from the investigation of optimization potential in the previous section 8.2.1, an optimal feature subset minimizing the number of input features while simultaneously maximizing the prediction performance in terms of accuracy is located at about 7 - 12 input features. Especially in this range, the new ERBE FS algorithm incorporating with LOO FS and GA FS the corresponding feature ranking and feature subset optimization principles clearly outperforms the other filter and wrapper methods RELIEF and LOO FS. Even more, the ERBE FS technique already reached an optimum in terms of the required number of input features and prediction accuracy at around eight input features.

Feature	No.	RELIEF						LOO						ERBE					
Flow_Argon_3	3									1	1	1	1						
Flow_Helium_1	7						1												
Pressure_Helium_1	9								1	1	1	1	1			1	1	1	1
Pressure_Helium_2	10							1	1	1	1	1	1	1	1	1	1	1	1
Pressure_Chamber_1	11		1	1	1	1	1			1	1	1	1		1	1	1	1	1
Temperature_Dome_1	13			1	1	1	1												1
Temperature_Dome_2	14					1	1												
Temperature_Dome_4	16		1	1	1	1	1				1	1	1						
Temperature_Dome_5	17				1	1	1												
Power_Chuck_1	21												1						
Voltage_Chuck_1	22									1	1	1	1			1	1	1	1
Flow_Helium_3	24			1	1	1	1												
Flow_Helium_4	25						1												
Flow_Helium_5	30												1						
Flow_Helium_8	33						1												
Counter_1	34																	1	1
Counter_10	43																		1
Counter_11	44																	1	1
Power_Bias_1	55			1	1	1	1						1				1	1	1
Power_Bias_3	57	1	1	1	1	1	1				1	1	1			1	1	1	1
Power_TS_1	60	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Power_TS_3	62	1	1	1	1	1	1		1	1	1	1	1						1
Power_TS_6	65											1	1				1	1	1
Power_Bias_6	67					1	1												
Flow_Silane_1	68											1	1	1	1	1	1	1	1
Logistic_1	73				1	1	1	1	1	1	1	1	1		1	1	1	1	1
Sum of features		3	5	8	10	12	15	3	5	8	10	12	15	3	5	8	10	12	15

Table 8.10: Feature composition of investigated RELIEF filter, LOO wrapper and ERBE FS algorithms for feature subset size 3, 5, 8, 10, 12 and 15. Every column of RELIEF, LOO and ERBE presents a feature subset where the sum of included features is stated at the bottom. Power_TS_1 is the only feature included within all feature subsets.

At last, table 8.10 lists the composition of the investigated feature subsets for RELIEF, LOO and ERBE FS. Only the eight features selected by the ERBE FS algorithm (cf. table 8.7) yielding the almost optimal prediction performance as illustrated in figure 8.29 are considered to be important since more features barely add significant information thereby improving the accuracy. The feature Power_TS_1 is the only one included in all feature subsets in the top three features and thus clearly identified by all approaches as indispensable input. The features Pressure_Helium_1 and Pressure_Helium_2 frequently selected only by LOO and ERBE FS are substituted by Flow_Helium_2 on recommendation of process experts (cf. section 8.2.1). Pressure_Chamber_1, Power_Bias_3 and Logistic_1 are recognized by all techniques as crucial whereas Voltage_Chuck_1 is again only considered by LOO and ERBE FS. Flow_Silane_1 is again only by ERBE FS revealed to be crucial and at least ranked at number 12 by LOO FS.

Summary: The results of the conducted experiments (cf. chapter 7) to evaluate the newly developed advanced VM system (cf. chapter 5) and the invented smart ERBE FS (cf. chapter 6) are illustrated and analyzed.

At first, prediction accuracy and reliability of the advanced VM system are tested on chronological ordered, unseen and independent data simulating a productive environment of several equipment processing a huge variety of products and basic types determined by a high diversity regarding other logistical parameters (e. g. recipe, operation). Data of two out of five equipment show major adjustments right at the transition from training to test resulting in unsatisfiable prediction performance. After retraining of new VM models including adjusted instances good results are achieved for $EQ1$ & $EQ2$ (cf. subsections 8.1.1 & 8.1.2). The predictions for equipment $EQ3$ – $EQ5$ (cf. subsections 8.1.3, 8.1.4 & 8.1.5) yield from the beginning good results even though the number of predictions outlying the control limits is in focus of investigation and further reduction. In overall, a reliability of mostly higher than 50 % of the detected outliers as well as a prediction accuracy in terms of CV(RMSE) of constantly less than 1 % clearly approves the feasibility and applicability of the investigated advanced VM system.

Secondly, the new ERBE FS algorithm (cf. section 6.5) is executed and the results of the various ERBE stages are evaluated in subsection 8.2.1. Here, the algorithm is tested to reveal the most important features for data of the HDP CVD process on equipment AMAT Centura (cf. section 8.2.1) and the PECVD process on equipment AMAT Producer (cf. section 8.2.1). In order to assess the optimization potential of the new technique, a manual feature subset optimization after execution of ERBE FS is conducted including process expertise and comprehensive knowledge of experienced process engineers. The results of the ERBE FS algorithm for both processes as well as the results from the investigated optimization potential clearly demonstrates the compelling capability of the new ERBE FS algorithm to tackle the problems and challenges regarding efficiency, scalability, knowledge discovery and accuracy (cf. subsection 4.1.1). Strong monotone increasing prediction performance in terms of accuracy measured by the outlier sensitive CV(RMSE) up to the optimum of the bias-variance-tradeoff (cf. section 3.3.4) and the discovery of the crucial feature subset is impressively achieved by the developed ERBE FS algorithm hence enabling an advanced VM system.

At third, the new ERBE FS technique is compared to the state of the art in terms of FS methods (cf. subsection 4.1.2) i. e. to the established ML filter technique RELIEF and the SVR-based wrapper technique LOO FS. The outlined results in subsection 8.2.2 explicitly corroborates the superior performance of ERBE FS for advanced VM especially in the crucial range for the size of the optimal feature subset (cf. Figure 8.29).

In the following chapter 9 a concluding discussion of the results for the developed smart FS algorithm enabling an advanced VM system is conducted considering the scope of the current state of the art and the related achievements regarding the fab-wide application in leading-edge SM industry.

9 Discussion

The obtained results of the experiments composed in the previous chapter to investigate the advanced VM system as well as the new ERBE FS algorithm are subsequently discussed in the next two sections.

In the first half of the present thesis covering a one and a half years time period the advanced VM system was projected and only limited resources could be assigned for development and implementation. The close cooperation with other partners (e. g. Fraunhofer-Gesellschaft) in the European research project IMPROVE implicates the challenge to work simultaneously on various other work packages (e. g. Predictive Maintenance) but also provides the opportunity to gain valuable knowledge in terms of ML algorithm development and inevitable DP for various environments. To the end of the IMPROVE project two core hypotheses are postulated by the author of the present thesis:

1. The choice of the ML induction algorithm used for prediction is less important as long as relevant boundary conditions are satisfied (e. g. good regularization ability).
2. Data preparation is crucial to optimize prediction performance and clearly outperforms the impact of the choice of the ML learning algorithm.

The firm conviction of the correctness of these hypotheses led to the two major aspects of the present work mainly investigated, developed, implemented and tested in the course of the second half of the three years time period at the Infineon frontend manufacturing site Regensburg:

1. Efficient development and implementation of an advanced VM system including the consideration, adjustment and enhancement of a suitable DM concept (CRISP-DM) incorporating all essential tasks needed for realization of a productive VM application:
   1.1. A VM system module (i. e. PTM) is implemented for the two scenarios of prediction of the defined target and training of the implemented ML algorithms (i. e. SVR, NN and M5').
   1.2. A VM module for configuration (i. e. CM) is developed to dynamically choose the best prediction model depending on logistical parameters and available data.
2. Data preparation consists of several tasks with FS as one of the most essential aiming on the reliable selection of an optimal subset of input variables. In addition to increasing the prediction performance of an induction technique, further advantages according to subsection 4.1.1 are realizable by smart FS:

2.1. A generic FS approach applicable for all suitable manufacturing processes maximizes efficiency of VM for corporate-wide fast implementation and deployment (cf. Efficiency).

2.2. The reduction of the feature subset not only decreases model complexity but also reduces corporate data storage, data traffic and computational effort and so improves scalability of an advanced VM system which can be assessed as a substantial monetary benefit to significantly strengthen the competitive advantage of Infineon in the SM industry (cf. Scalability).

2.3. The revelation of the most important process parameters by knowledge discovery enables further process developments in future which is highly desirable to extend the potential to gain competitive advantage in SM (cf. Knowledge Discovery).

2.4. Incorporating only the most important process parameters and with it only valuable and curcial information maximizes reliability and prediction performance by means of highest accuracy in terms of CV(RMSE) (cf. Accuracy).

9.1 Advanced Virtual Metrology System

A comprehensive assessment of the advanced VM system includes a CBA to estimate the expected benefit resulting from the productive application to the dedicated HDP CVD process, the adapted CRISP-DM approach as well as the VM development and implementation at Infineon. The DM approach can hardly be tested itself and is therefore discussed in terms of mastering the projected challenge. The deployed advanced VM system is tested on five productive equipment and the obtained results outlined in section 8.1 are assessed in detail together with the implementation of the Prediction and Training Module as well as the Configuration Module. Finally, the comparison with current state of the art VM system completes the discussion.

Cost Benefit Analysis: The CBA constituted in subsection 5.1.1 and appendix A.3 is filled with required economic and productive data in order to calculate costs and benefits related to the implementation of the advanced VM system and its application to the HDP CVD process. Costs only arise for deployed human resources to develop and implement the VM system application itself. For the calculation of the benefits the affected process areas HDP CVD, chemical mechanical planarization and metrology are distinguished. The benefits of reduced metrology and improved cycle time (cf. equations A.3 and A.4, respectively) are calculated for the first two process areas whereof the latter also profits by a reduction of produced scrap wafers (cf. equation A.5). The benefits for metrology are an improved utilization (cf. equation A.6) of the already installed measurement equipment resulting in savings by reduced metrology operations (cf. equation A.7). The potential and enormous benefit of avoiding new purchase of expensive metrology equipment can hardly be expressed as a monetary benefit at that time as it mainly depends on future and actually unknown production ramp-up strategies triggered by the global market development for SM. As final result of the CBA calculation the full amortization of the

VM development will take place three years after productive application of VM only for the HDP CVD process. Deploying the advanced VM system by contemporary rollout of the current implementation to other processes and process areas (e. g. physical vapor deposition, plasma etch or chemical mechanical planarization) will substantially accelerate the attainment of the break-even point as the related costs are significantly reduced compared to the initial development at a quite similar level of achievable benefits. After the break-even point the developed advanced VM system will annually generate substantial benefits and strengthen the competitive advantage of Infineon in the SM industry.

Knowledge Discovery and Data Mining for Virtual Metrology: For the development and implementation of the advanced VM system, the CRISP-DM approach is chosen, adapted and elaborated for all comprised phases. Thereby, a well-structured DM procedure is conducted within a short period of time of only one and a half year whereas indeed, the first phase (i. e. Business Understanding) is mainly performed during the first half of the present work within the European research project IMPROVE. The extensive literature research as well as an insightful exchange with other researchers in the field of VM led to the previously postulated hypotheses as ultimate but still in its completeness unsolved challenges (cf. subsection 4.1.1). A high focus on the iterative CRISP-DM core phases (cf. figure 3.1) paired with a strict concentration on the essential problems enabled the efficient development of a first VM system. A final evaluation is performed on a subsequent and thereby independent DS and the results outlined in section 8.1 clearly demonstrate the applicability of the advanced VM system for efficient fab-wide deployment whereat even in case of major equipment maintenance activities (cf. EQ_1 and EQ_2) a solution is found to achieve good results. In addition to approved efficiency, scalability by reduction of model complexity affecting data storage, data traffic and computational effort is accomplished. Knowledge discovery to enhance future process developments is highlighted in the subsequent assessment and discussion of smart FS. Finally, the optimization of the prediction performance in terms of accuracy and reliability as mandatory requirement for VM in SM is also demonstrated by the conducted experiments.

In the end, all relevant phases are successfully executed and the hereinafter discussed results in addition to recent publications [89], [90] corroborate the approach of adapting CRISP-DM for VM in SM.

9.1.1 Assessment of Results

General Observations & Accuracy: The result of the primary predictions based on the complete and MW DS for equipment EQ_1 and EQ_2 are not satisfying and related to the illustrated manual adjustments of some of the most important process parameters. The secondary prediction for EQ_1 and EQ_2 on the updated MW DS and all predictions for EQ_3 – EQ_5 yield a RMSE $\leq 1\,\%$ relative deviation of the LT. While the VM models for EQ_1 and EQ_3 are predestinated subjects of future investigations and improvements, the predictions based on updated MW data for EQ_2 and both complete and MW data for EQ_4 & EQ_5 are unconditionally acceptable for productive application. In general, the results achieve the required objective of highest prediction performance. The MAE is given as an additional measure for reference, but regarding the objective to precisely predict outliers the more sensitive RMSE is primarily considered.

In general, the variance of the prediction delivered by the final VM models tend to exceed the variance of the observed target for all investigated equipment $EQ_1 - EQ_5$. This is mainly related to the chosen reduced logistical granularity in order to build only one generic VM model based on this logistic but valid for various other logistical granularities which could be designated to the specific equipment, process chambers and recipes. The alternative to build, evaluate, run and maintain specific VM models subdivided into much more logistical granularities is generally not feasible in low-volume-high-mixture SM due to the enormous variety of logistical variation. An exception could be a dedicated VM model in very specific cases for exceptionally critical products (e. g. decreased process stability in combination with increased process requirements).

The sensitivity to detect outliers by the VM models despite increased variance yields good results for accurate predictions (i. e. based on updated MW data for EQ_2 and both complete and MW data for EQ_4 & EQ_5) discovering most of the outliers for the investigated equipment.

The prediction performance for the MW approach outperforms the VM models trained on the corresponding complete DS in any case with the exception of the primary predictions for EQ_1 and EQ_2. In this regard, not the slightly lower prediction error achieved by the MW approach but the considerably smaller size of the used training DS is remarkable.

Model Fit & Coefficient of Determination: In general the results point out that due to the negative outcome of R^2 no exclusive assessment of the model fit can be achieved by using R^2. The complex HDP CVD process (cf. section 2.3) can obviously neither be assumed to be explained nor to be fitted by a linear function corroborating the investigation of the applied RBF kernel (cf. equation (6.1)). Nevertheless, some correlation of R^2 with the accuracy of the model fit is visible as for an increasing coefficient of determination from negative values towards zero the prediction error perceptibly decreases (cf. evaluation of prediction for EQ_1 and EQ_2 in subsections 8.1.1 and 8.1.2, respectively). From the comparison of the results for all equipment, it can be derived that a $R^2 >$-3 indicates an overall good prediction.

Outliers & Sensitivity: Even though the reliability of the prediction models is estimated quite well by the calculated sensitivity, shortcomings are observed in case of DSs comprising only very few outliers due to excellent process control and the missing possibility to generate additional outliers in a productive environment by dedicated design of experiments. The occurrence of only few outliers exacerbates the challenge to develop a VM system due to the lack of especially valuable information within training data. Thus, exhaustive outlier detection by any VM model cannot be guaranteed whereas the already established process control is still significantly improved by the established all-over Wafer-to-Wafer control.

The integration of VM into statistical process control and so the inclusion of VM predictions into the calculation of the UCL and LCL may yield expanded control limits implying a more or less degraded centered process capability index which assesses the overall process stability. As long as an increasing variance introduced by the prediction compared to the observed metrology (cf. EQ_4 & EQ_5 in subsections 8.1.5 & 8.1.4) does not significantly broadens these *control* limits (e. g. less than 1 % as visible for EQ_4 & EQ_5), no further action is required because highest accuracy and quality are still ensured due to significantly broader process *specification* limits. Well-defined *specification* limits and insignificantly expanded *control* limits could come in handy regarding the effect of previously discussed and currently present small logistical granularity (i. e. only few considered logistics within the configuration or as input variables). The majority of falsely predicted outliers (e. g. EQ_4) would then be located within the slightly broadened *control* limits thus facilitating the application of VM demonstrating a concurrent benefit and solution in case of smaller logistical granularity. For a significant higher variance or shift of the prediction mean versus the observed mean (e. g. EQ_1 & EQ_2) and thus an unacceptably centered process capability index value, a second enhancement in addition to the MW approach is possible. The incorporation of further but not much more logistical parameters (e. g. only product) either for configuration by logistical granularity or as input variables would yield further improvement in terms of accuracy and precision due to additional information and characteristics inherent in these logistics even though resulting in increased complexity of the advanced VM system. Undetected outliers close to or even outside the *specification* limits can be revealed since the very sensitive SVR model shows already good sensitivity in terms of outlier detection around the *control* limits. Furthermore, the results of EQ_1 and EQ_2 approve the high prediction performance of the SVR method itself where significantly degraded accuracy is recognized very well if major shifts or single noticeable outliers are present in any of the crucial features (cf. Temperature_Dome_3 and Voltage_Chuck_1 for EQ_1 and the latter also for EQ_2).

The achieved prediction performance in terms of accuracy and reliability unconditionally meets the high demands on VM made in SM (cf. Accuracy) and clearly demonstrate the applicability and capability of the advanced VM system.

Training on MW or complete DS: Due to periodical equipment maintenance activities including major adjustments of crucial process parameters (e. g. Voltage_Chuck_1) right at the transition from training to test DSs for EQ_1 and EQ_2, the initial prediction does not yield satisfying results for neither the MW nor the complete DS approach. However, these most important features with values showing a significant offset are valuable and cause the VM model to predict outliers or an entire offset. If a continuous offset occurs the retraining of the VM model with mainly most recent instances for the updated MW approach is indicated. With an increasing number of instances after periodical equipment maintenance the VM model can be further optimized if required in case of still unsatisfying results caused by too less new instances within the training DS. The application of the newly introduced RI with traffic light logic (cf. section 5.2.2) compares the prediction with outcomes of various other ML techniques to exclude the degradation of a VM model only based on SVR. Furthermore, the similarity index according to [25] performs an early warning to not rely on the VM prediction if a significantly offset is observed in the value of a crucial feature. The updated MW approach successfully demonstrated the capability and applicability of an advanced VM system to rapidly achieve good prediction performance within a changing environment indicated by coefficient of determination and sensitivity and clearly corroborated by the CV(RMSE) for EQ_1 and EQ_2 (cf. figure 8.10).

All suitable VM models built on MW DSs for $EQ_3 - EQ_5$ achieved lower deviation from the target (w. r. t. CV(RMSE)), higher model fit approximated by R^2 and comparable or higher number of detected outliers than those trained on the complete DS. Thus, an adequately sized MW approach appears to be favorable. At last, the influence of the variety of manufactured products included in a DS over time is reduced in a MW DS due to the fact that the generally positive effect of an increased amount of training instances comes along with a higher variation of the feature values and thus a more generalized VM model.

9.1.2 Implementation

The design and implementation of the generic advanced VM system (cf. subsection 5.2.2) is developed and realized within a time period of two years and tested for applicability for the complex HDP CVD process considering all related systems and data flows (cf. subsection 5.1.2). The model training quadratically depends on the number of used instances and linearly on the number of features which are minimized by smart FS and thus not decisive in a range of less than ~20 input features out of originally far more available variables. The training of new VM models is performed within several minutes and in parallel to online VM prediction in the internal framework thereby not delaying the production since the processing of an entire lot takes place within 0.5 - 2 hours and including some additional minutes for reloading the equipment a newly trained VM model is available. Once more, the MW DS approach is favorable in terms of computational time effort as well as required memory size due to smaller DSs needed to train a new VM model. The entire sequence to compute all VM predictions online for an entire lot is performed at most in 10 s with an average of ~5 s whereat the step which dominates the execution time is determined by DB I/O processes to load and store data or VM models. The prediction itself for the entire lot (i. e. for all 25 wafers) takes place within a fraction of a second.

Hence, the performance of new advanced VM system is also corroborated with respect to the investigated challenges of efficiency and scalability (cf. subsection 4.1.1) and demonstrating the capability of future application and deployment in other process areas.

Virtual Metrology Prediction and Training Module

As highlighted in section 5.2.2 and illustrated in figure 5.2, the configured PTM trains the required VM models with regard to the necessary logistical granularity and predicts the metrology outcome based on provided data. A basic VM approach using only SVR as prediction technique including crucial DP was already prototyped and described in an earlier publication [90]. Additionally, comparable ML methods in terms of prediction accuracy are implemented for the advanced VM system (cf. figure 5.3) to improve the reliability of the prediction [93]. Thus, the already stated RI in section 5.2.2 can serve as estimation of the prediction quality and will be evaluated in future in more detailed studies.

Virtual Metrology Configuration Module

The CM (cf. section 5.2.2) is developed to efficiently handle various logistical granularities occurring due to the combinations of the enormous amount of all the available logistical parameters (e. g. product, basic type, process group, operation, recipe, equipment). The implementation of various ML techniques for VM is investigated in recent research for many different logistical granularities separating data into more dedicated DSs to achieve highest accuracy required for VM in SM (cf. section 4.2). Considering the fact that the application of higher granularity implying the inclusion of more logistical parameters into the DS could yield more accurate predictions due to an incorporation of more specific characteristics within available data, it is remarkable that the developed VM system achieves high accuracy even for a logistical granularity configured by the CM based on only three logistical parameters and only a single logistic added to the input feature set. Maintenance of higher detailed granularity with logistical parameters comprising a high number of different values (e. g. basic type) turns out to be infeasible due to the resulting highly fragmented DSs not containing enough data to reasonably train an induction method. The challenge of high data fragmentation in low-volume-high-mixture SM is accepted and mastered by incorporating only four logistical parameters for the entire VM system.

9.1.3 Comparison with current State of the Art Virtual Metrology

The fundamental challenges (cf. subsection 4.1.1) to sustainably deploy VM in SM are corporate-wide efficient applicability and deployment without enormous additional effort for investigation of each process area (Efficiency), scalable enterprise systems with minimized data storage, data traffic and computational effort (Scalability), the ability to reveal only the crucial process parameters containing all essential information (Knowledge Discovery) and high prediction performance in terms of accuracy and reliability (Accuracy).

The new advanced VM system introduced in chapter 5 and discussed in detail in the previous subsections remarkably demonstrates the capability to master all these challenges. Especially the difficult task of concrete knowledge discovery to enable and enhance future process developments is achieved by smart FS and extensively highlighted in the next section. The new advanced VM system is subsequently compared to the actual state of the art research provided in chapter 4 for FS for VM (cf. section 4.1) and VM (cf. section 4.2).

Comparison with FS for VM: As first challenge the efficiency of the new advanced VM system enabled by smart FS is approved by development and subsequent application of the new ERBE FS algorithm for the different processes HDP CVD and PECVD performed on the different production equipment AMAT Centura and AMAT Producer (cf. section 2.4). A lack of efficient deployment of FS for VM according to section 4.1 is observed in *NN for CVD*, *NIPALS for CVD* which also lacks reliability in contrast to acceptable accuracy, *GA FS for Etch* with improvable reliability as well as *FS and Projection for Etch* which shows only partially acceptable accuracy as tradeoff for remarkable dimension reduction and scalability.

Scalability as second challenge is achieved by the new ERBE FS algorithm by significant dimensionality reduction of 85 % of the input feature set from initially 75 features (omitting the artificial ones) down to 11 features in the final subset. In contrast a lack of scalability is identified in *Tree Ensemble for Etch*, *Canonical Analysis for PVD* with additional potential for improvement regarding efficiency, *Clustering for Etch* with also insufficient capability in terms of knowledge discovery and *Aggregative Linear Regression for Etch* again not meeting the requirements of knowledge discovery and efficiency.

A comparison of FS methods for VM and the new ERBE FS algorithm in terms of knowledge discovery including the revelation of crucial features is highlighted in more detail in subsection 9.2.3.

The results presented in section 8.1 clearly demonstrate the ability of the new advanced VM system enabled by smart FS to achieve highest accuracy and reliability as required in SM industry. In contrast, a lack of accuracy is observed in current research as *SVR for Yield* where scalability is also subject to further improvements and *Recursive Coefficient Centering for Critical Dimension* which shows insufficient ability to tackle any of the four stated problems. In terms of mature reliability of outlier detection as well as efficiency, scalability and knowledge discovery *SVM for Outlier Detection* yields unsatisfying results regarding FS for VM.

Comparison with VM: In addition to FS for VM, the entire new VM system with the combination of the PTM and the CM as described in section 5.2.2 and discussed above successfully demonstrates the ability of a productive VM implementation in the SM industry to master all the challenges according to subsection 4.1.1 in comparison with various other VM systems.

While a fab-wide equipment monitoring and FDC system (cf. *TSMC*) yields a high coefficient of determination, the problems of efficiency and scalability are not tackled so far. *Forward Selection Component Analysis for Etch* achieves only moderate accuracy, neglects the importance of scalability and also struggles to provide a complete VM system to efficiently transfer and deploy VM applications to other process areas.

A lack of scalability, efficiency and knowledge discovery is observed for the VM systems outlined in *NN for Chemical Mechanical Planarization* and *Wafer-fine R2R Control* whereas the a partially missing quantified assessment of the latter hinders an extensive comparison to other VM approaches.

Since more than a decade the extensive research of *Cheng et al.* addresses a wide range of problems with VM in SM but even though noticeable results are presented for this variety of problems the challenge of knowledge discovery is not considered so far.

Finally, unstated accuracy together with a lack of scalability, knowledge discovery and missing evaluation of efficiency are drawbacks of $\mathcal{L}_1$*-penalized ML for CVD*.

9.2 Smart Feature Selection

The development of the ERBE FS algorithm is motivated by the goal to combine the merits of various FS approaches under the condition that highest prediction accuracy as well as precision and reliability are crucial to deploy VM as generic application in a productive SM environment.

Subsequently the results of the new ERBE technique (cf. section 6.5) are discussed followed by its comparison to the established RELIEF filter method (cf. section 4.3) and the LOO FS wrapper approach (cf. section 6.3), finally complemented by a comparison to the current state of the art.

9.2.1 ERBE Feature Selection

The evaluation of the developed ERBE FS algorithm is performed on two different production equipment, AMAT Centura & AMAT Producer, on which different manufacturing processes, HDP CVD & PECVD respectively, are running. The more detailed investigation of the new ERBE FS technique on the first type of equipment contains the pure execution of the FS algorithm as well as a manual optimization to estimate how close the result of the automated ERBE FS method comes to an optimal solution potentially obtained by further improvement. The affirmation of ERBE as a generic FS algorithm is effectuated by an automated execution for a different and independent production equipment and process.

ERBE FS for AMAT Centura

ERBE FS: The execution time of the ERBE FS algorithm strongly depends on the size of the DS as well as the system on which it is run and the applied algorithm settings (cf. appendix A.5). The exhaustive FS on a DS containing 2699 instances lasts in total more than one week due to the quadratic dependence of the computational time on the number of instances. The time-determining step of SVR training is accelerated by optimizing the calculation of the H matrix (cf. subsection 6.5.3) but is still time consuming for DSs with more than around 1000 instances. Adjustments of the ERBE FS parameters (e. g. reduction rate) are also thinkable to accelerate the processing. Since the execution of the ERBE FS algorithm is performed only once to reveal the crucial features to improve all future prediction models and can be executed in parallel to other applications a longer computational time appears to be tolerable (cf. appendix A.5).

The scalability depends on the number of features (i. e. incorporated process parameters) and thereby the required data storage, data traffic and computational effort necessary to run the advanced VM system. Although, only 75 features remained after strict DP for the HDP CVD process the new ERBE FS algorithm achieved a reduction of 85 % to 11 features left in the feature subset after ERBE stage 8. Hence, the remarkable dimensionality reduction enables a high performance VM system and tackles the problem of scalability.

The ERBE FS curve for all ERBE stages in figure 8.26 plots the prediction accuracy as relative deviation from the target (i. e. LT) against the number of features including the artificial ones representing the model complexity and so the scalability of the entire VM system. Minor improvements and degradations are observable within part I and II (i. e. ERBE stages 1 - 6) which are expected to be caused by statistical variation of the different training/validation DS compilation related to shuffling of the instances as well as the randomly varying artificial variables. Part III demonstrates the potential to successfully differentiate features by efficient LOO FS with less computation effort compared to GA FS.

An equivalent accuracy compared to the best result of the ERBE FS algorithm obtained at ERBE stage 3 with 56 input features is achieved for several considerably smaller DSs (cf. figure 8.27). Thus, the hypothesis of a single feature or feature subset combination causing this improvement can be rejected. A small degradation of the prediction performance can be observed for remaining 24 and 16 input features whereas a clear degradation of the accuracy is visible for only 8 input features (i. e. only 3 real features excluding the 5 artificial ones). The impressive prediction accuracy by SVR demonstrates highest accuracy with already 75 input features and the scalability is constantly improved.

At last, the demand for FS for VM in SM is observed since the reduction of feature subsets to their minimum still achieving highest prediction performance minimizes the risk of degradation of the VM model influenced by random effects totally incoherent to the physical process itself (e. g. interferences of magnetic fields with a cable with degraded isolation by other tools causing the measured maximum value to shoot up). If many dispensable features are included in the feature subset an extreme outlier in any of these features contributing mainly noise may mislead the VM model yielding a false prediction since the behavior of this feature was distributed within an inconspicuous range of values so far. Furthermore, it is observed for equipment 1 and 2 in subsections 8.1.1 and 8.1.2 that the explicit change (e.g. offset out of the range of training data) of only one crucial feature can already significantly degrade the prediction performance of a VM model.

Investigation of Optimization Potential: The resulting feature subset *ERBE_FS_S8_11* (represented by a burnt orange diamond in figure 8.27) composed of 11 features of the 8^{th} ERBE stage (excluding the 5 artificial variables) yields a remarkable low error evaluated by the RMSE of 0.572 %. The hypothesis to achieve highest prediction accuracy by feature subset *ERBE_FS_S8_11* cannot be rejected due to the fact that within the other 15 manually composed feature subsets the best score is achieved for feature subset *FS_OPT_14_2* with a relative deviation from the target LT of 0.537 %. The minor improvement of the prediction accuracy of 0.035 % is expected to be caused by statistical variation of the different training/validation DS compilation. Hence, it is clearly corroborated that the new ERBE FS algorithm developed in the present thesis found a solution for the required highest prediction accuracy.

The feature subset *FS_ERBE_S8_11* automatically computed by the new ERBE FS algorithm achieved the same high prediction accuracy as the manually optimized feature subset *FS_OPT_8_2* which is due to its reduced size used for the evaluation of the advanced VM system. Table 9.1 lists an extraction of the feature compilation of both subsets for better comparison.

As it is already highlighted in section 8.2.1, the two features Pressure_Helium_1 and Pressure_Helium_2 are substituted by Flow_Helium_2 and the feature Temperature_Dome_3 is additionally incorporated into the feature subset *FS_OPT*_8_2 based on the recommendation of the process experts. Visualized in table 9.1, these differences between the feature subsets *FS_OPT_8_2* and *FS_ERBE_S8_11* are four out of in total seven differences. According to section 2.3, the substitution of the process parameters Pressure_Helium_1 and

Feature	Number	FS_OPT_8_2	FS_ERBE_S8_11
Flow_Helium_2	8	1	
Pressure_Helium_1	9		1
Pressure_Helium_2	10		1
Pressure_Chamber_1	11	1	1
Temperature_Dome_3	15	1	
Voltage_Chuck_1	22	1	1
Counter_11	44		1
Power_Bias_1	55		1
Power_Bias_3	57	1	1
Power_TS_1	60	1	1
Power_TS_6	65		1
Flow_Silane_1	68	1	1
Logistic_1	73	1	1
Sum of features		8	11

Table 9.1: Feature subset comparison of *FS_OPT_8_2* and *FS_ERBE_S8_11*: Columns 1 & 2 state the features with their numbers and column 3 lists the manually optimized feature subset *FS_OPT*_8_2. The last column provides the feature subset composition of stage 8 of the ERBE FS algorithm excluding artificial features resulting in 11 features. The total sum of feature is added at the bottom.

Pressure_Helium_2 by Flow_Helium_2 is reasonable since all three variables are related to the quantity of helium flowing between the wafer backside and the electrostatic chuck whereat either the helium pressure within this isolated space or the injected helium volume per time is measured, respectively. The helium gas is responsible for the adjustment of the wafer temperature by cooling the backside of the wafer during the HDP CVD process. Thus, the high impact on the deposition target (i. e. LT) is directly caused by each of these three highly interrelated features due to the fact that the deposition of SiO_2 at the wafer surface strongly depends on the temperature (cf. section 2.3). For this reason the temperature of the process chamber assessed at the dome (cf. figure 2.3) by the feature *Temperature_Dome*_3 was recommended by the process experts to be included in *FS_OPT*_8_2 assuming to add by this important information to feature subset *FS_OPT*_7 relevant for the process outcome, but almost identical results in terms of accuracy for *FS_OPT*_8_2 and *FS_OPT*_7 (the latter even slightly better - cf. figure 8.27) rejected this hypothesis whereas the capability of ERBE FS in terms of knowledge discovery is further approved.

Due to the fact that the variable Counter_11 is not included neither in any other feature subset of the expert selection or manual optimization (cf. table 8.6) nor selected by another FS algorithm (cf. table 8.9) but only in the feature subset *FS_ERBE_S8_11* of ERBE stage 8 and not physically interrelated to other process parameters of the HDP CVD process, the hypothesis of a last remaining insignificant feature might be corroborated.

The last two different features Power_Bias_1 and Power_TS_6 only present in feature subset *FS_ERBE_S8_11* are most probably highly correlated with Power_Bias_3 and Power_TS_1, respectively, since the two other FS algorithms (i. e. RELIEF filter & LOO wrapper) also selected these features as the most important ones as observed in table 8.9.

The common feature subset of *FS_OPT_8_2* and *FS_ERBE_S8_11* is composed of the following six features whose description with respect to the HDP CVD process is provided for further understanding of the importance of these process parameters:

1. **Pressure_Chamber_1**: Pressure within process chamber
2. **Voltage_Chuck_1**: Voltage applied to electrostatic chuck
3. **Power_Bias_3**: DC-bias power applied by RF coil generator
4. **Power_TS_1**: Top/Side power applied by RF coil generator
5. **Flow_Silane_1**: Silane gas flow into process chamber
6. **Logistic_1**: Logistical Parameter

As Counter_11 is suspect to be the only feature within the resulting feature subset not contributing valuable information it is remarkable that disregarding the variables Counter_11 and the two ulteriorly correlated features Power_Bias_1 and Power_TS_6, the new ERBE FS algorithm revealed the best and most likely optimal feature subset containing 8 out of the 75 original features with only process parameters absolutely indispensable and required to achieve highest prediction performance for VM in SM.

Furthermore, the results obtained for the ES feature subsets composed of 20 and 22 variables (cf. figure 8.27) are clearly inferior compared to the ERBE FS results. The unexpected and significantly higher error of ~ 0.7 % and ~ 0.8 % compared to the entire ERBE FS except the last stage raises the question about the root cause for this significant difference. An investigation of the feature subsets according to table 8.6 yields the result that a single feature is missing in the two ES feature subsets which is never eliminated by the ERBE FS algorithm and resulted as one of the three remaining most important features: *Power_TS_1*. Hence, in terms of Knowledge Discovery in Databases an important finding is primarily obtained in the present thesis: The newly developed ERBE FS algorithm originally published in [94] revealed a feature classified as irrelevant by experienced process experts to be indispensable and absolutely crucial to be included into the feature subset of a VM prediction model to achieve highest accuracy for VM in SM.

ERBE FS for AMAT Producer

In contrast to ERBE FS performed on data of the HDP CVD process on the equipment AMAT Centura, the present execution is performed within less than two days with 826 instances but an almost tripled number of features (i. e. 198 vs. 75 input variables) in the DS. The only linear dependency on the number of input features (in contrast to the quadratic subjection to the input instances) demonstrates the feasibility and applicability of the ERBE FS approach even for higher dimensionality of the feature space. Thus, the case of a highly populated input space with more than 10.000 variables according to [125] is negligible compared to only ~ 100 - 200 additional instances. Furthermore, an increased reduction rate during part I of the ERBE FS algorithm can yield an adequate solution for huge input dimensionalities.

Up to now the SM industry is still struggling to efficiently deploy a corporate-wide VM system also revealing only the crucial process parameters and neglecting noisy information due to the lack of any generic FS technique (cf. section 4.1.1). The successful application of the new ERBE FS algorithm to the PECVD process performed on the equipment AMAT Producer (cf. section 2.4) clearly approves the efficiency of the developed advanced VM system enabled by smart FS to master the economic challenges in highly competitive SM.

The prediction accuracy vs. model complexity for the PECVD process as illustrated in figure 8.28 corroborates the capability of ERBE FS to significantly reduce the number of features while constantly increasing the prediction performance. From originally 198 input process parameters the ERBE FS algorithm eliminated more than 92 % of these variables to yield a remarkable small feature subset containing only 15 features (excluding the artificial ones). Thus, the problem of scalability for a corporate-wide VM system with the challenge to efficiently handle data storage, data traffic and computational effort for all VM prediction models is successfully tackled and approved.

In contrast to FS for the HDP CVD process no degradation of the prediction accuracy proving the loss of crucial information for a VM model is observed after the last ERBE stage. A minimum of the error would indicate an optimum and perfect bias-variance-tradeoff of the model complexity curve (cf. figure 3.2). The increase of the error for an insufficient number of incorporated features is not yet observed and thus it cannot be assumed that an optimal completely purified feature subset is achieved after the last ERBE stage. Therefore, the optimal number of process parameters is expected to be in a similar range as for the HDP CVD process around 5 up to the 15 revealed features. A fine tuning feature optimization of the ERBE FS technique with adjusted parameters (e. g. smaller reduction rate for ERBE part III) appears to be appropriate for initial DSs with high number of features yielding more detailed knowledge discovery. Thus, the optimum of the model complexity curve at which the prediction degrades would finally become visible again if an adjusted setting for the ERBE FS algorithm with increased focus on fine-tuning optimization is applied.

The initial poor prediction performance with a relative deviation from the target LT of 1.8 % is insufficient in terms of highly accurate and reliable predictions required for VM. However, the constant improvement of the accuracy down to ~ 1.1 % again approves the capability of an advanced VM system to optimize the prediction performance by smart FS. A further decrease of the error excluding the five artificial features from the final feature subset can be assumed as it is observed for the ERBE FS for AMAT Centura whereas an even higher improvement is most likely for AMAT Producer since the accuracy is not yet obviously minimized as for AMAT Centura yielding 0.05 % improvement. Hence, it can be deduced that the new ERBE FS algorithm finally enables VM for the PECVD process similar to HDP CVD without the need of manual optimization to achieve a sufficiently small error and by this approving the entire concept of generic VM in SM with highest accuracy.

The strong monotonic decrease of the error in figure 8.28 of the different ERBE stages indicates a significant performance degradation of an ML induction algorithm in the presence of too many noisy and interfering features. It is clearly observed that the decrease slows down after ERBE part I and increases again after ERBE stage 7. Conducting the LOO FS the first significant reduction of the RMSE in the first three ERBE stages in part I eliminates the extremely noisy and interfering process parameters. The features removed during the next four ERBE stages do obviously neither contain relevant information nor disturb the prediction model. Due to the fact that the first ERBE stage of part III (i. e. ERBE stage 7) also yields only a minor decrease of the error the hypothesis of only small improvements of the accuracy are caused by GA FS in part II can be rejected. The approach to provide an initial faster feature reduction by applying LOO FS compared to an optimization performed by GA FS (cf. equation 6.3) is clearly validated by the results of the first three ERBE stages. Also a faster fine tuning feature optimization in the last ERBE part III by LOO FS compared to GA FS is corroborated by the significant improvement of the prediction performance.

The assessment of the new ERBE FS algorithm for data of the HDP CVD and PECVD process performed on the equipment AMAT Centura and AMAT Producer, respectively, yields several insightful aspects. A higher range of the reduced error caused by a significantly larger initial feature set containing noisy and inferring features and the constant improvement of the prediction performance are obtained for the PECVD process. High accuracy already from the beginning as well as the final degradation of the VM model are observed for the HDP CVD process. A combination of both yields an idealized model complexity curve optimizing the regularized risk as introduced in the bias-variance-tradeoff (cf. figure 3.2). Hence, the applicability of the new ERBE FS algorithm is demonstrated for both scenarios of the model complexity: high bias/low variance and low bias/high variance.

In addition to the present detailed discussion of the execution of the entire ERBE FS algorithm for the AMAT Producer, the results of the individual ERBE stages are assessed (in the same way as previously for the AMAT Centura) in the appendix on page xxvii together with an exhaustive feature list on page xxiii.

9.2.2 Assessment of ERBE, RELIEF and single LOO Feature Selection

A comparison of the ERBE FS algorithm with the well-established RELIEF FS filter and the LOO FS wrapper method is outlined in subsection 8.2.2 where feature subsets comprising 3, 5, 8, 10, 12 and 15 features are investigated and the top ranked features are listed in table 8.9. Figure 8.29 corroborates the superior performance of the new ERBE FS algorithm especially in the crucial range around 7 - 12 features.

Both ERBE FS and LOO FS preserve the process parameters Pressure_Helium_2 (related to the wafer temperature as discussed before in section 9.2.1) and Power_TS_1 as two of only three input features for the VM prediction model. The only difference of selecting the feature Flow_Silane_1 by ERBE FS instead of Logistic_1 by LOO FS yields a remarkable improvement of the prediction accuracy of ~ 0.6 % as the error of 1.51 % is reduced to 0.93 % for ERBE FS. Thus, the crucial inherent information of the feature **Flow_Silane_1** is directly and as fast as possible revealed by the new ERBE FS algorithm and subsequently approved by the significantly decreased CV(RMSE) achieved by LOO FS for feature subset of size 12 also containing even this process parameter. The two most important process parameters selected by ERBE FS (i. e. Pressure_Helium_2 & Flow_Silane_1) are completely neglected by RELIEF FS.

In contrast to these feature subsets of size 3, the RELIEF FS filter technique selected only Power_TS_1 as common variable and Power_TS_3 and Power_Bias_3 as other input variables resulting in a quite good relative prediction error CV(RMSE) of only 1.08 %. Hence, the inclusion of these features can partially compensate the missing intrinsic information of Flow_Silane_1 and Pressure_Helium_2. Especially the feature **Power_Bias_3** corroborates the fact of more inherent characteristics since it is also included within the ERBE FS feature subset containing 8 process parameters achieving a noticeable smaller error than the comparable subset of the LOO FS method with also 8 variables whereat the immediate inclusion of Power_Bias_3 and Temperature_Dome_4 in the following feature subset of LOO FS composing 10 variables also yields higher accuracy. Hence, the absolutely indispensable process parameter **Power_TS_1** is obviously the most important feature immediately selected by all algorithms.

The logistical parameter Logistic_1 and the process parameter Pressure_Chamber_1 are subsequently identified as most important by ERBE FS yielding the so far best result of 0.74 % for 5 input features. While the feature Pressure_Chamber_1 is almost concurrently selected by RELIEF FS it is considered by LOO FS for the first time within the feature subset of size 8. Similarly but with a little higher difference between the FS techniques, Logistic_1 is ranked by LOO FS already as 3^{rd} and by RELIEF FS as 9^{th} most important feature. Thus, logistical parameter **Logistic_1** and the process parameter **Pressure_Chamber_1** clearly contribute crucial information to enable further improvement of the prediction performance.

While Power_Bias_3 is already stated as one of the additional 3 process parameters for the next ERBE FS feature subset of size 8, Voltage_Chuck_1 and Pressure_Helium_1 are also incorporated by ERBE FS as well as selected and ranked by LOO FS (as number 7 and 4, respectively) but neglected by RELIEF FS. Hence, **Pressure_Helium_1** and **Pressure_Helium_2** are immediately identified as crucial features by ERBE FS and LOO FS contributing indispensable characteristics to assess the wafer temperature and by this to finally achieve highest prediction performance. In order to include the same intrinsic information, these two variables are substituted by Flow_Helium_2 as recommended by the process experts yielding firstly remarkable results as demonstrated for unseen data (cf. section 8.1) and secondly further reduction of the dimensionality of the feature set. **Voltage_Chuck_1** as eighth process parameters contained in feature subset of size 8 selected by ERBE FS also serves as important source of information and indicator of potential data shift as corroborated in figure 8.4.

Compared to LOO FS and ERBE FS the fast degradation of the results (cf. figure 8.29) obtained by the feature subsets $\geq$ size 10 as well as the significantly differing composition of top ranked features neglecting crucial process parameters demonstrated the deficiency of the RELIEF FS filter method despite of its computational advantage to reliably perform dimensionality reduction for VM.

The results of LOO FS and ERBE FS in table 8.9 yield a quite similar feature subset composition but noticeably inferior prediction performance of LOO FS as illustrated in figure 8.29. Hence and according to the principle "A good feature ranking criterion is not necessarily a good feature subset ranking criterion." [49], the comparison of both FS methods underlines the superior performance of the new ERBE FS algorithm and thus the importance of the incorporation of a heuristic feature subset selection technique like the introduced GA FS.

The precisely revealed optimum of absolutely indispensable features to achieve highest prediction accuracy is clearly demonstrated by the new ERBE FS algorithm in figure 8.29 yielding an almost perfectly minimized relative deviation from the target LT of 0.64 % with the feature subset containing eight features highlighted above and outlined in table 8.9. Even more, the new ERBE FS algorithm (cf. section 6.5) remarkably approved its specific ability to discover new process knowledge by immediate detection of Power_TS_1 as most important feature. Hence, the new ERBE FS algorithm enables an advanced VM system for corporate-wide VM in SM.

9.2.3 Comparison with current State of the Art FS Methods

The challenge of knowledge discovery (cf. section 4.1.1) for VM to reveal crucial logistical and process parameters as support for process experts for future developments is investigated and encountered by various research and already approved by the new ERBE FS algorithm (cf. discovery of feature Power_TS_1 in section 9.2.1). While several state of the art FS methods are developed for VM in SM (cf. subsection 4.1.2), a lack of knowledge discovery is still observed in *Canonical Analysis for PVD* with more potential for improvement of scalability and efficiency for generic VM deployment, *PLS for PECVD*, *Tree Ensemble for Etch* with additional shortcomings in terms of scalability for a corporate-wide VM system and *Clustering for Etch*.

The results presented in section 8.2 clearly demonstrate the ability of the new ERBE FS algorithm to yield highest prediction performance as required in SM industry. In contrast, insufficient accuracy is achieved by *SVR for Yield* also struggling to attain scalability as well as by *Recursive Coefficient Centering for Critical Dimension* and *SVM for Outlier Detection* where the latter two approaches are unable to tackle the problems of scalability and efficiency.

Even though the computational effort for the previously discussed FS methods is not comprehensively provided for comparison, the execution time of the new ERBE FS algorithm can be considered as quite long ranging from a couple of hours to several days depending on the amount of input features and mainly instances. As an advantage compared to investigated state of the art FS, the new ERBE FS algorithm focusses on outlier detection by incorporation of the sensitivity due to the use of the RMSE as intrinsic evaluation criterion to assess the prediction accuracy. Compared to the most described FS methods for VM (cf. subsection 4.1.2), the new ERBE FS technique is also able to prevent local optima by inclusion of a heuristic search (i. e. GA FS) which is independent of the applied kernel function and thus provides a solution for shortcomings of other FS methods summarized in section 4.3.2. Another advantage is given by the ability of the new ERBE FS method to deal with multiple selection criteria in terms of concurrent optimization of the number of input features and prediction accuracy overcoming the drawbacks recapitulated in section 4.3.1. Furthermore, the effective distinction between noisy variables and features containing valuable information is improved by the introduction of artificial features. The variety of adjustable parameters of the ERBE FS algorithm (e. g. definition of percentage for each reduction phase, continuous reduction vs. fine tuning in later stages, SVR model parameters, definition of thresholds for transition between ERBE parts, number and distribution of artificial features, etc.) also enables a high level of adaption to manifold processes and use cases beneficial for generic deployment of VM in SM.

The challenge to find a suitable FS method to unify the benefits of computational efficient feature ranking techniques for fast feature elimination and heuristic feature subset optimization constantly incorporating crucial interdependencies is successfully mastered by the new ERBE FS algorithm by combining the merits of LOO FS and GA FS with the principles of structural and empirical risk minimization. The remarkable results corroborate the capability of the new ERBE FS technique to overcome the so far unresolved problems of efficiency, scalability, knowledge discovery and accuracy and thus to enable an advanced VM system in the very demanding and complex SM industry.

Summary: The results outlined in the previous chapter 8 are discussed in detail to assess the developed advanced VM system and the new ERBE FS algorithm.

The economic requirement in the highly competitive SM industry for effective development and efficient deployment of a generic VM system is initially verified by a CBA and subsequently fulfilled by application of the CRISP-DM approach. The evaluated results of the advanced VM system (cf. subsection 9.1.1) are approved by a remarkable accuracy achieved to minimize the deviation from the target, a good model fit to the observed data and the noticeable reliability to detect outliers. Prediction models are build and approved on the complete and the favored MW DSs. The reviewed implementation of the two VM modules (cf. subsection 9.1.2) for prediction and training as well as configuration (i. e. PTM & CM) demonstrates the corporate-wide applicability of the VM system. A final comparison with the current state of the art in VM (cf. subsection 9.1.3) corroborates the solution of the so far unresolved problems regarding efficiency, scalability, knowledge discovery and accuracy by the advanced VM system.

The imperative demand for VM in SM to implement scalable applications in terms of data storage, data traffic and computational effort by reduction of the high number of possible logistical and process parameters is obviously recognized but so far not successfully mastered by suitable FS techniques. The new ERBE FS algorithm overcomes the problems of scalability and knowledge discovery while guaranteeing highest prediction performance and by this enables an advanced VM system for efficient deployment in SM. The successful application of the new ERBE FS technique is approved for different processes (i. e. HDP CVD & PECVD) performed on different production equipment (i. e. AMAT Centura & AMAT Producer). An investigation of the optimization potential of the result obtained by the new ERBE FS method corroborates an excellent selection of an almost perfect feature subset only containing the really crucial features to achieve highest prediction accuracy. Here, a by process experts supposed important process parameter is identified not to improve the prediction performance. Even more, a so far neglected process parameter is discovered as most important feature contributing valuable information for future process development and enhancement (cf. subsection 9.2.1). The new ERBE FS algorithm outperforms an established FS filter technique as well as FS wrapper method and approves the potential to reveal only the crucial features to meet the challenging requirements of VM (cf. subsection 9.2.2). Again, a final comparison with the current state of the art in FS for VM (cf. subsection 9.2.3) corroborates the mastery of the so far unresolved problems efficiency, scalability, knowledge discovery and accuracy by the new ERBE FS algorithm thus providing a solution for the scientific challenge to enable FS for VM.

10 Conclusion and Outlook

A final conclusion regarding the development and implementation of an advanced Virtual Metrology (VM) system enabled by smart Feature Selection (FS) realized with the new Evolutionary Repetitive Backward Elimination (ERBE) FS algorithm is drawn in the following. The concept of VM in high-mixture-low-volume Semiconductor Manufacturing (SM) is approved by ensuring efficiency, scalability, knowledge discovery and accuracy required for corporate-wide productive application. Finally, an outlook is provided for possible future research and development as well as further beneficial enhancements of the current application.

10.1 Conclusion

Advanced Virtual Metrology System

The introduced Cost-Benefit Analysis (CBA) enabled the economic assessment of the development and implementation of VM within the industrial environment of productive SM at the Infineon Technologies AG. The planning of future VM enhancements (e. g. similarity index - cf. subsection 4.1.2) and deployments can be based on the available CBA calculations to estimate an expected return on invest.

The effective implementation paired with constantly focusing on crucial research activities to ensure a very tight schedule is assured by the well-structured CRoss Industrial Standard Process for Data Mining (CRISP-DM) procedure. Similar to professional project management, firstly specific targets are defined and regularly aligned with the stakeholders, secondly the iterative execution of the core phases steadily improves the entire VM system while permanently controlling the progress and thirdly after evaluation and acceptance by the stakeholders the corporate-wide VM system is finally deployed.

The VM system is successfully evaluated by Root Mean Squared Error (RMSE) for accuracy, sensitivity for reliability and Coefficient of Determination (R^2) for model fit and is proven to generate good prediction models built on Moving Window (MW) and complete Datasets (DS). Noticeable reliability to detect outliers, sufficiently good model fit to the observed data and remarkable accuracy minimizing the deviation from the target are achieved for various equipment and processes.

The developed Prediction and Training Module (PTM) and Configuration Module (CM) demonstrate the ability to perform automated predictions and configurations for a productive and corporate-wide VM system in SM. Only four logistical parameters are sufficient for the High Density Plasma (HDP) Chemical Vapor Deposition (CVD) and Plasma Enhanced Chemical Vapor Deposition (PECVD) processes to achieve the projected goals whereas depending on data availability and specifically required prediction model accuracy more logistics can be included in the logistical granularity scenario managed by the generic CM or as additional input feature in a DS. Different high-sophisticated learning methods (i. e. Neural Network (NN), Decision Tree M5' (M5') and Support Vector Regression (SVR)) are successfully implemented to be executed in parallel for further improvement of the prediction reliability and to design a new Reliance Index (RI) comparing the outcome of these Machine Learning (ML) methods [93].

The observed computing performance of some seconds to execute online VM predictions including data storage and data traffic corroborates the feasibility of the advanced VM system enabled by smart FS.

The expectations in terms of VM reducing physical metrology to a feasible minimum for assuring a self-controlled VM system are met. The results of the prediction performance are not satisfying if major equipment maintenance is recently performed which in fact is immediately identified via comparison with physical metrology and the future RI. Thus, the investigated MW approach based on a smaller and more recent DS yield an approach to tackle the problem to maintain highest prediction performance.

A final comparison with the current state of the art in VM corroborates the solution of the so far unresolved problem of simultaneously mastering the challenges efficiency, scalability, knowledge discovery and accuracy by the advanced VM system.

Finally, the results of similar performing ML algorithms combined with the improvements by means of smart FS confirm the two stated core hypotheses that not the choice of the induction algorithm significantly improves the accuracy, but the reduction to a feature subset containing only the most important features.

Smart Feature Selection

The development of the new ERBE FS algorithm enfolds the incorporation of three different objectives by consecutive parts. In part I, fast dimensionality reduction initially eliminates noisy and inferring features by Leave-One-Out (LOO) FS. In part II, feature subsets are optimized keeping crucial interdependencies and preventing solutions at local optima by Genetic Algorithm (GA) FS. In part III, feature fine tuning optimization is conducted by LOO FS to reveal only the most important features composing the final feature subset.

The introduction of artificial features perceptibly supports the discrimination between important and dispensable features and serves as an effective threshold used as transition criteria to move to the next part of the ERBE FS algorithm.

The new ERBE FS technique approved its powerfulness to reduce the initial feature set by 85 % and 92 % for different processes (i. e. HDP CVD & PECVD) performed on different production equipment (i. e. Applied Materials (AMAT) Centura & AMAT Producer) while yielding a remarkable highest prediction accuracy of 0.57 %.

An investigation of the optimization potential of the results obtained by the new ERBE FS method corroborates the excellent selection of an almost perfect feature subset only containing the really crucial features to achieve highest prediction accuracy. Hence, the new ERBE FS algorithm enables scalability of the advanced VM system for corporate-wide deployment.

The total computational effort of some hours up to some days spent for execution of the new ERBE FS technique is acceptable for manageable DSs. Compared to the iterative process of Data Mining (DM) consuming some weeks up to several months the process of FS is performed only once if a purified DS is available and can also be scheduled in parallel as an independent task.

The comparison of the new ERBE FS algorithm with the Expert Selection (ES) demonstrates the potential of knowledge discovery for VM in SM. A feature so far neglected and classified as redundant by the process experts is discovered to be one of the most important features contributing valuable information for future process development and enhancement.

Furthermore, established and high-sophisticated FS techniques (i. e. RELIEF filter and SVR-based LOO wrapper) are compared to and outperformed by the new ERBE FS method approving its potential to rapidly reveal only the crucial features while meeting the outlined four challenging requirements of VM.

A final comparison with the current state of the art in FS for VM corroborates the mastering of the so far unresolved problem regarding simultaneous achievement of the challenges efficiency, scalability, knowledge discovery and accuracy by the new ERBE FS algorithm thus providing a versatile solution for the scientific challenge of enabling FS for VM.

Summary: In the end, the imperative demand for VM to efficiently implement scalable applications in terms of data storage, data traffic and computational effort by reduction of the high number of possible logistical and process parameters yielding highest prediction performance is obviously recognized in the research area of VM but so far not adequately satisfied by available FS methodologies. For the first time, the efficient development and application of a FS technique successfully meeting all these challenges is proven in the present work for different fabrication processes performed on different production equipment. The new ERBE FS algorithm masters the challenges of *scalability* and *knowledge discovery* while guaranteeing highest prediction *accuracy* and by this enables an advanced VM system yielding high *efficiency* by fastest deployment in SM.

10.2 Outlook

The final objective to approve the concept of smart FS to enable advanced VM at Infineon is realized in a short period of time. So, several additional challenges and options for extension and enhancement as well as further investigation emerged in the course of the present thesis. The following outlook outlines some of the remaining aspects related to possible future work.

Advanced Virtual Metrology System

- The training of ML models on MW and complete DSs yield good but not always perfect prediction results. The strategy of retraining ML models based on differently sized and dynamic DSs can be optimized especially for the case of prediction performance degradation due to shifts of input parameters related to major maintenance activities.

- An investigation of various scenarios with different criteria (e. g. applicability of the mentioned RI introducing a traffic light logic, statistical significance tests with respect to the outlier distribution) as best point in time to trigger a new training of the VM prediction model to achieve best accuracy and reliability is already content of actual research continued at Infineon by the author of the present thesis.

- Up to now the target for the VM prediction to monitor the HDP CVD process is the thickness of the deposited dielectric layer. Other existing process control parameters (e. g. the associated refraction index and the bow of the processed wafer resulting in mechanical stress of the deposited layer) are also monitored by physical metrology and thus could be subject to future VM application.

Smart Feature Selection

- Regarding the noticeable computational effort of the ERBE FS technique, a future incorporation of a correlation analysis between the repetitive ERBE stages appears to be conceivable to further speed up the feature subset selection.

- Recent research focused on the optimization of the applied kernels with mixed polynomial, hyperbolic or Radial Basis Function (RBF) kernels. Moreover the usage of different kernels for each feature depending on their characteristics appears to be worthy of further investigation.

- The most important features are already subject to detailed discussions with process engineers whereas the crucial feature newly revealed by the ERBE FS algorithm needs further profound investigation involving extensive expertise in the area of unit process development at Infineon to fully exploit the potential for future enhancements of the deposition process and development of new processes based on the gained knowledge.

A Appendix

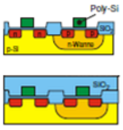
A.1 Frontend Process Areas

In today's SM, circular silicon wafers with highest purity $> 99.9999\,\%$ of various sizes with a typical diameter of 150 mm, 200 mm or 300 mm are used in frontend fabs whereof 200 mm sized wafers are commonly organized in lots containing 25 identical wafers. These lots are processed in all frontend process areas multiple times to build up a layered structure onto the wafers. In general, SM frontend process areas can be organized into six process sequences for altering the physical structure of a wafer. Additionally, two process sequences (i. e. Clean & Metrology) either prepare wafers for the next one or to control the result of the last one. Figure 2.1 outlines the non-modifying process sequences metrology and clean (center) and the altering process sequences layer composition, planarization, structuring, layer removal, layer transformation and resist strip [174], [66].

The overall process flow is to build up a layer in layer composition, followed by planarization of the new layer, structuring of the plane layer, removal and/or transformation of the structured layer and resist strip of the structuring. Between any of these process sequences, the wafers can be cleaned and required measurements can be performed. The different process areas included in these process sequences are briefly highlighted below.

Layer Composition

Layer composition can be divided into the two process areas deposition and furnace which can be further broken down according to the reaction type in terms of chemical or physical deposition and chemical layer composition or oxidation in furnace processes.

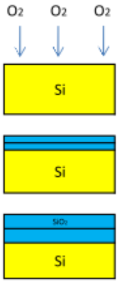


Deposition

Deposition

Chemical deposition is used in SM industry for the deposition of thin films of solids onto substrates (i. e. wafers) by chemical reaction of a certain mixture of process gases within a process chamber. Physical deposition includes evaporation deposition of various materials as well as sputtering of designated materials to deposit the sputtered elements.

To initiate the deposition process as well as to increase the deposition rate, the reaction gases can be activated thermally, electrically (by plasma), chemically or by photons. The properties of the deposited conductive or dielectric film are determined by the method of activation, the applied energy, the amount and chemical properties of the supplied gases, as well as the temperature and the material properties of the substrate [174].

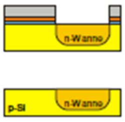


Oxidation

Furnace

To oxidate an already existing substrate on the wafer surface (e. g. Si to SiO_2), thermal oxidation takes place in an environment of high temperature (≈ 1000°C) as dry or wet oxidation using oxygen or water as oxidant, respectively, yielding different layer thicknesses which determine quality and cost of the oxidation process [42]. Conductive path isolation, masking for diffusion processes and protection from damage are some of the manifold functions of the various oxide layers [79]. Chemical layer composition in furnace processes differ from those discussed above in application of higher temperatures enabling additional chemical reactions yielding layers with different properties.

Planarization

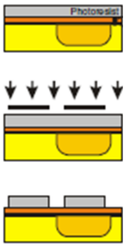


Chemical Mechanical Planarization

Chemical Mechanical Planarization

In order to obtain a plane wafer surface required for subsequent structuring processes, the wafer rotates head-down in the opposite direction to an also rotating polishing pad together with a slurry in between as polishing agent [42].

Structuring

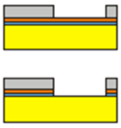


Lithography

Lithography

The size and shape of patterns mapped by lithography onto the wafer surface define the structure of the currently processed layer and mask regions for subsequent processing. Smaller sizes enable higher densities of electronic circuits integrated into an IC responsible for the computing capacity or memory size of chips [174]. After coating the wafer surface with a light-sensitive photoresist, the patterns for conductive paths or contact holes are constituted by the exposure to light emitted through structured photomasks followed by development of the dried patterns. For positive photoresists the developed regions on the wafer surface are dissolved, whereas negative photoresists are handled vice versa [79]. In the end, the wafer is baked out to solidify the remaining photoresist in order to protect the underlying structures in subsequent processes [162].

Layer Removal



Etch

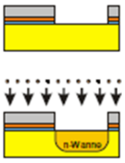
Etch

In SM, the intended regions of the masked patterns from previous lithography are removed from the wafer layer surface by either wet or dry etching holes or trenches into the material depending on the desired characteristics of these methods [162]. Chemical wet etching is characterized by liquid etching reagents like acid mixtures or undiluted acids with high purity [66].

Dry etching is divided into chemical etching (plasma etching) on the one hand where reactive gases (e. g. Cl_2) are streamed in the process chamber over the substrate and react with the material at the surface and physical etching on the other hand where ion bombardment (cf. section 2.3: sputtering) is used to excavate material out of the intended substrate [162].

Layer Transformation

Layer transformation is divided into annealing and doping which is further grouped into ion implantation and diffusion as processes to contaminate the crystalline silicon substrate with elements containing one valence electron more (e. g. N/P) or less (e. g. B) than tetravalent silicon. The former modification enables n-type conductivity where free electrons can transfer the charge and the latter modifications enables p-type conductivity where the charge is transferred by a hole in the crystal lattice [36], [79].



Ion Implantation

Ion Implantation

Dedicated elements for n-type or p-type conductivity are ionized in the plasma within the arc chamber of the implanter equipment. The positively charged ions are accelerated by an electrical field from the ion source through the mass resolving magnet and apertures selecting the specified ions which are further accelerated towards to the wafer surface. Due to the more or less amorphous structure of the substrate related to the impact of the implanted ions, a subsequent annealing process is necessary [42], [162], [36].

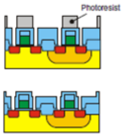
Annealing

Annealing describes the process of electrical activation of implanted elements by "healing" the atomic structure at high temperatures (≈ 1000°C) to arrange these elements at the atomic sites corresponding to the initial crystalline structure of silicon [36].

Diffusion

An inert carrier gas enriched with the element to be doped into the substrate is streamed through a quartz tube along the wafer surface at very high temperature. The concentration gradient causes a uniform diffusion of the dopant into the silicon crystal [79], [162].

Resist Strip



Resist Strip

Resist Strip

Depending on the processes performed between lithography and resist strip in terms of the applied temperatures affecting the durability of the remaining photoresist, the techniques for resist strip varies from application of dissolvers over wet etch processes to plasma ashing [66].

Clean

All of the aforementioned processes might contaminate the wafer surface with remaining chemicals or particles. Thus, the wafer needs to be cleaned between the various process steps by high-purity water, mixtures of chemicals, brushing, compressed inert gas or in ultrasonic baths [66], [162].

A.2 Benefit and Data Quality

As already outlined (cf. section 2.2), even for only regarding product cycle time reduction an enormous benefit of up to 10 % additional production volume is expected for fab-wide application of VM [22]. To make the right choice in an economic sense in terms of which metrology method and approach to use, more than basic cost and revenue factors like costs of purchase, installation and maintenance as well as tool footprint, necessary material and labor should be considered according to [152]. Primarily, product quality and related revenue associated with each metrology method are important aspects to be aware of together with the current market situation and actual quality control policies. In this regard, in-situ metrology as enabled by VM outperforms inline and offline metrology especially during the phase of new manufacturing processes whereas the obtained knowledge gain should be comparable. As final conclusion, during production ramp-up phases in-situ and inline measurements provide better response times than offline metrology [152].

Another economic assessment of VM in SM evaluates first potential risks of VM according to failure causes, modes and effects which could be a decrease of equipment uptime, an increase of production costs and scrap of productive wafers or lots. As second statement, economic benefits are distinguished by process equipment types and various potential savings are outlined. Reduction of out-of-control production as most valuable return on invest dominates savings in terms of metrology steps, shorter cycle time and higher equipment utilization. At last, the break-even is calculated for the implementation of VM for a plasma etch tool to be achieved after six quarters [84].

The aspect of data quality as already mentioned in 2.2 is investigated in more detail in [58] for VM in SM with CVD as an application. The problem within SM with inappropriateness and instability of data collection is well known. Poor prediction performance may occur due to incorrect, asynchronous and fragmented data which raises the attention to spend effort on data quality to achieve accuracy in addition to precision. Particularly, data reduction, data normalization, data cleaning and data anomaly detection are in focus of this work with a proven performance improvement of VM [58].

A major overview and assessment of VM implementation in SM is given in [18] for a 300 mm SM foundry in the CVD process area. Motivated by an expected and significant benefit, the aim to monitor tool performance and to detect quality of productive wafers in real-time in important SM areas (e. g. lithography, etch, deposition, planarization) justifies a considerable effort in research and development. In SM foundries the possible lack of the ability to control manufactured devices as final product due to unknown target functionality withhold by the customer further motivates the implementation of VM. At the beginning, a reasonable and in SM established formula to calculate the profitability of a SM fab is evaluated with focus on improvements by VM [18]:

$$Profitability = \sum_i \left(\frac{W_i * T_i * Y_i}{S_i * (P_i - C_i)} \right) - (MW * F) \tag{A.1}$$

with the number of monthly turn ratio:

$$T = \frac{720\, hours}{t_{process} + t_{metrology} + t_{wait}} \tag{A.2}$$

W: Number of wafer starts/month
T: Number of monthly turn ratio
Y: fab yield
P: Average selling price/wafer
S: Number of total stages of each product from wafer start to wafer out
C: Manufacturing cost/wafer
MW: Monitoring wafer cost for all tool monitoring processes
F: Tool monitoring frequency/month
i: Product index
t_{main}: Main processing time/wafer/stage/month
t_{metro}: Metrology operation time/wafer/stage/month
t_{wait}: Wait time in processing/wafer/stage/month

Table A.1: Profitability key factors for SM fab [18]

The individual key factors in equation (A.2) are explained in table A.1. Obviously, a reduction of $t_{metrology}$ in the denominator of equation (A.2) directly results in an increased profitability whereas Y, MW and F (i. e. not MW in this context) are also affected indirectly. For the present example, the basic assumption is made that the amount of test wafers used for metrology measurement equals 15 %–30 % of the daily production output. In result, a profit of 4 M. $ is calculated for 30.000 wafer starts per month in the SM foundry. Quite a few requirements for the implementation of a sustainable VM system are collected which describe the challenging approach to integrate a VM application into the IT infrastructure of a productive SM fab. Furthermore, the necessary effort for developing a framework to run VM comprising various data services (e. g. metrology data management, equipment data acquisition, computer integrated manufacturing, sensor data preprocessing) is underlined. In addition to already discussed benefits like reduction of physical metrology and real time tool control capability in APC, a VM application according to [18] provides an option to enhance equipment maintenance from reactive to predictive based on real-time metrology forecast, hence enabling process engineers to trigger necessary tool maintenance just in time. Also, validation of sensor data from the equipment which is performed by sensor preprocessing, a dedicated design of experiment and controller development modules to control the VM model development, can be enriched by VM due to the fact that even tool sensors does not always reliably provide accurate data for prediction models. The already emphasized data quality becomes a crucial aspect for VM implementation because the "Garbage-in Garbage-out" scenario does not only result in incorrect metrology forecasts, but also cause productivity and yield loss. Hence, a VM system to validate sensor data is desirable. As first conclusion pursuant to [18], VM is proven to assure process quality and tool performance and to improve overall equipment effectiveness while reducing physical metrology for tool monitoring. As final conclusion and corroboration of the motivation, VM is expected to evolve to a standard operation and the necessary VM framework design will be a key component of computer integrated manufacturing [18].

A.3 Cost-Benefit Analysis

As stated in various references (cf. chapter 4), the implementation of a fab-wide VM system is expected not only to improve quality but also to yield a significant economic benefit. Hence and to supplement the VM system at Infineon, the economic benefits of the VM implementation are assessed in the present work for the use case of HDP CVD at the frontend manufacturing site in Regensburg.

As a result of the corporation with the Fraunhofer-Gesellschaft within the European research project IMPROVE involving the already referenced research paper [84] (cf. section A.2), adjustments and enhancements of the basic concept for investment assessment were developed. Future VM rollout activities can be planned and performed based on the subsequently highlighted evaluation formalism for CBA to effectively promote VM implementation for use cases with the highest expected benefit. In the following, the CBA is discussed as far as permitted with regard to the eligible protection of Infineon's confidentiality interests.

A.3.1 Costs

Nowadays, a variety of cost elements is related to development projects within industrial economy (e. g. invest for IT hardware, software and licenses, service, support and internal personnel costs as well as capital costs). For only the development of the VM system including the ERBE algorithm, only personnel costs need to be considered in the present thesis. The emerged costs within the different organizational areas of Infineon were added up to obtain the associated total employees full time equivalent for the final break-even estimation (cf. chapter 8).

A.3.2 Benefits

The annual benefits are split into three parts which cover the processes affected by the investigated HDP CVD use case for VM implementation.

Benefits for High Density Plasma Chemical Vapor Deposition: Starting with immediate benefits, the total reduction of metrology costs for HDP CVD – $RedMetro$ – was calculated according to equation (A.3). Here, the wafer starts per month – $WaferStarts$, the months – M, the process steps – $Steps$, the sampling rate – $Sampling$, the measured wafers – $WafersMeas$, the real metrology reduction factor achieved by VM – $RedFac$ and the metrology cost of ownership – CoO were multiplied and divided by the wafers measured per carrier – $WafersPerCarrier$:

$$\mathbf{RedMetro} = \frac{WaferStarts * M * Steps * Sampling * WafersMeas * RedFac * CoO}{WafersPerCarrier} \tag{A.3}$$

Furthermore, the cycle time improvement due to less real metrology and thus faster production flow was estimated and a monetary benefit assigned according to equation (A.4) based on the following variables: Metrology time – $MetroTime$, sampling rate – $Sampling$, process steps –

Steps, reduction factor – *RedFac*, wafer starts per month – $WaferStarts$, Averaged value of wafer – $WaferVal$ and interest rate – *IntRate*:

$$\textbf{ImpCycleTime} = MetroTime{*}Sampling{*}Steps{*}RedFac{*}WaferStarts{*}WaferVal{*}IntRate \tag{A.4}$$

Benefits for Chemical Mechanical Planarization: The benefit of avoiding scrap wafers in the subsequent chemical mechanical polishing process due to increased stability of the HDP CVD process related to the VM implementation was assessed as reduction of scrap – *RedScrap* according to equation (A.5). Again, a product of following input variables was calculated: number of events resulting in scrap wafers – $NumEvents$, months – M, number of affected equipment – $NumEQ$, number of scrap wafers per event – $NumScrapWafer$, average value of product wafer – $ProdWaferVal$, reduction factor – $RedFac$:

$$\textbf{RedScrap} = NumEvents{*}M{*}NumEQ{*}NumScrapWafer{*}ProdWaferVal{*}RedFac \tag{A.5}$$

Benefits for Metrology: By the reduction of physical measurements due to application of VM more capacity of the metrology equipment will be available for other measurements. Hence, on the one hand, this improved utilization enables higher production output if physical metrology represents a bottleneck operation and on the other hand as a second major advantage it helps to avoid the purchase of new and usually expensive metrology equipment during a capacity ramp-up. The first equation (A.6) calculates the improved utilization – *ImpUtil* from the product of the total HDP wafers per month – $HDPWafers$, the rate of the measured HDP wafers within a lot – $RateMeasWafersInLot$ and the reduction factor – $RedFac$ divided by the total number of measured wafers per month – $MeasWafersPerMonth$:

$$\textbf{ImpUtil} = \frac{HDPWafers * RateMeasWafersInLot * RedFac}{MeasWafersPerMonth} \tag{A.6}$$

The resulting economic benefit in terms of only higher production output results in case of physical metrology is operated as bottleneck can be derived according to equation (A.7) as metrology savings – $MetroSavings$ from the fraction of the product of the total number of measured wafers per month – $MeasWafersPerMonth$, the months – M, the average cost of a measurement – $AvgMeasCost$ and the previously calculated improved utilization – *ImpUtil* as numerator and the total number of measured wafers per lot $MeasWafersPerLot$ – as denominator:

$$\textbf{MetroSavings} = \frac{MeasWafersPerMonth * M * AvgMeasCost * ImpUtil}{MeasWafersPerLot} \tag{A.7}$$

The expected significant benefit of postponing the purchase of new and expensive metrology equipment cannot be calculated with sufficient reliability as it mainly depends on future and actually unknown production ramp-up strategies.

A.4 Feature Overview

A.4.1 AMAT Centura

Index	Feature	Category	Unit	Description
1	Flow_Argon_1	Gas flow	sccm	Argon gas flow into process chamber
2	Flow_Argon_2	Gas flow	sccm	Argon gas flow into process chamber
3	Flow_Argon_3	Gas flow	sccm	Argon gas flow into process chamber
4	Flow_Argon_4	Gas flow	sccm	Argon gas flow into process chamber
5	Temperature_Chamber_1	Temperature	°C	Temperature within process chamber
6	Temperature_Chamber_2	Temperature	°C	Temperature within process chamber
7	Flow_Helium_1	Gas flow	sccm	Helium gas flow into process chamber
8	Flow_Helium_2	Gas flow	sccm	Helium gas flow into process chamber
9	Pressure_Helium_1	Pressure	mTorr	Helium pressure at electrostatic chuck
10	Pressure_Helium_2	Pressure	mTorr	Helium pressure at electrostatic chuck
11	Pressure_Chamber_1	Pressure	mTorr	Pressure within process chamber
12	Pressure_Chamber_2	Pressure	mTorr	Pressure within process chamber
13	Temperature_Dome_1	Temperature	°C	Temperature at ceramic dome
14	Temperature_Dome_2	Temperature	°C	Temperature at ceramic dome
15	Temperature_Dome_3	Temperature	°C	Temperature at ceramic dome
16	Temperature_Dome_4	Temperature	°C	Temperature at ceramic dome
17	Temperature_Dome_5	Temperature	°C	Temperature at ceramic dome
18	Current_Chuck_1	Current	mA	Current applied to electrostatic chuck
19	Current_Chuck_2	Current	mA	Current applied to electrostatic chuck
20	Current_Chuck_3	Current	mA	Current applied to electrostatic chuck
21	Power_Chuck_1	Power	W	Power applied to electrostatic chuck
22	Voltage_Chuck_1	Voltage	V	Voltage applied to electrostatic chuck
23	Voltage_Chuck_2	Voltage	V	Voltage applied to electrostatic chuck
24	Flow_Helium_3	Gas flow	sccm	Helium gas flow into process chamber
25	Flow_Helium_4	Gas flow	sccm	Helium gas flow into process chamber
26	Load_1	Loading	-	Equipment loading check
27	Load_2	Loading	-	Equipment loading check
28	Flow_Oxygen_1	Gas flow	sccm	Oxygen gas flow into process chamber
29	Flow_Oxygen_2	Gas flow	sccm	Oxygen gas flow into process chamber
30	Flow_Helium_5	Gas flow	sccm	Helium gas flow into process chamber
31	Flow_Helium_6	Gas flow	sccm	Helium gas flow into process chamber

Feature Overview (1) of available process and logistical parameters for AMAT Centura including artificial features

Index	Feature	Category	Unit	Description
32	Flow_Helium_7	Gas flow	sccm	Helium gas flow into process chamber
33	Flow_Helium_8	Gas flow	sccm	Helium gas flow into process chamber
34	Counter_1	Counter	-	Counter at equipment tool parts
35	Counter_2	Counter	-	Counter at equipment tool parts
36	Counter_3	Counter	-	Counter at equipment tool parts
37	Counter_4	Counter	-	Counter at equipment tool parts
38	Counter_5	Counter	-	Counter at equipment tool parts
39	Counter_6	Counter	-	Counter at equipment tool parts
40	Counter_7	Counter	-	Counter at equipment tool parts
41	Counter_8	Counter	-	Counter at equipment tool parts
42	Counter_9	Counter	-	Counter at equipment tool parts
43	Counter_10	Counter	-	Counter at equipment tool parts
44	Counter_11	Counter	-	Counter at equipment tool parts
45	Counter_12	Counter	-	Counter at equipment tool parts
46	Counter_13	Counter	-	Counter at equipment tool parts
47	Counter_14	Counter	-	Counter at equipment tool parts
48	Counter_15	Counter	-	Counter at equipment tool parts
49	Counter_16	Counter	-	Counter at equipment tool parts
50	Counter_17	Counter	-	Counter at equipment tool parts
51	Counter_18	Counter	-	Counter at equipment tool parts
52	Counter_19	Counter	-	Counter at equipment tool parts
53	Counter_20	Counter	-	Counter at equipment tool parts
54	Counter_21	Counter	-	Counter at equipment tool parts
55	Power_Bias_1	Power	W	DC-bias power applied by RF coil generator
56	Power_Bias_2	Power	W	DC-bias power applied by RF coil generator
57	Power_Bias_3	Power	W	DC-bias power applied by RF coil generator
58	Power_Bias_4	Power	W	DC-bias power applied by RF coil generator
59	Power_Bias_5	Power	W	DC-bias power applied by RF coil generator
60	Power_TS_1	Power	W	Top/Side power applied by RF coil generator
61	Power_TS_2	Power	W	Top/Side power applied by RF coil generator
62	Power_TS_3	Power	W	Top/Side power applied by RF coil generator
63	Power_TS_4	Power	W	Top/Side power applied by RF coil generator
64	Power_TS_5	Power	W	Top/Side power applied by RF coil generator
65	Power_TS_6	Power	W	Top/Side power applied by RF coil generator
66	Power_TS_7	Power	W	Top/Side power applied by RF coil generator
67	Power_Bias_6	Power	W	DC-bias power applied by RF coil generator
68	Flow_Silane_1	Gas flow	sccm	Silane gas flow into process chamber
69	Flow_Silane_2	Gas flow	sccm	Silane gas flow into process chamber
70	Flow_Silane_3	Gas flow	sccm	Silane gas flow into process chamber
71	Flow_Silane_4	Gas flow	sccm	Silane gas flow into process chamber
72	Counter_22	Counter	-	Counter at equipment tool parts
73	Logistic_1	Logistics	-	Logistical Parameter
74	Tune_1	Tuning	-	Equipment tuning check
75	Tune_2	Tuning	-	Equipment tuning check
76	Artificial_1_ran	Artificial	-	Gaussian artificial feature
77	Artificial_2_ran	Artificial	-	Gaussian artificial feature
78	Artificial_3_ran	Artificial	-	Uniform artificial feature
79	Artificial_4_cor	Artificial	-	Gaussian duplicated artificial feature
80	Artificial_5_cor	Artificial	-	Uniform duplicated artificial feature

Feature Overview (2) of available process and logistical parameters for AMAT Centura including artificial features

A.4.2 AMAT Producer

Index	Feature	Category
1	Flow_Argon_1	Gas flow
2	Flow_Argon_2	Gas flow
3	Flow_Argon_3	Gas flow
4	Flow_Argon_4	Gas flow
5	Flow_Argon_5	Gas flow
6	Flow_Argon_6	Gas flow
7	Flow_Argon_7	Gas flow
8	Pressure_Chamber_1	Pressure
9	Pressure_Chamber_2	Pressure
10	Pressure_Chamber_3	Pressure
11	Pressure_Chamber_4	Pressure
12	Pressure_Chamber_5	Pressure
13	Pressure_Chamber_6	Pressure
14	Pressure_Chamber_7	Pressure
15	Pressure_Chamber_8	Pressure
16	Pressure_Chamber_9	Pressure
17	Pressure_Chamber_10	Pressure
18	Pressure_Chamber_11	Pressure
19	Pressure_Chamber_12	Pressure
20	Pressure_Chamber_13	Pressure
21	Pressure_Chamber_14	Pressure
22	Power_Bias_1	Power
23	Power_Bias_2	Power
24	Power_Bias_3	Power
25	Power_Bias_4	Power
26	Power_Bias_5	Power
27	Power_Bias_6	Power
28	Power_Bias_7	Power
29	Power_Bias_8	Power
30	Logistic_1	Logistics
31	Counter_1	Counter
32	Power_Heater_1	Power
33	Power_Heater_2	Power
34	Power_Heater_3	Power
35	Power_Heater_4	Power
36	Power_Heater_5	Power
37	Power_Heater_6	Power
38	Power_Heater_7	Power
39	Power_Heater_8	Power
40	Power_Heater_9	Power
41	Temperature_Heater_1	Temperature
42	Temperature_Heater_2	Temperature
43	Temperature_Heater_3	Temperature
44	Temperature_Heater_4	Temperature
45	Temperature_Heater_5	Temperature
46	Temperature_Heater_6	Temperature
47	Temperature_Heater_7	Temperature
48	Temperature_Heater_8	Temperature
49	Temperature_Heater_9	Temperature
50	Temperature_Heater_10	Temperature
51	Temperature_Heater_11	Temperature
52	Temperature_Heater_12	Temperature
53	Temperature_Heater_13	Temperature
54	Temperature_Heater_14	Temperature
55	Temperature_Heater_15	Temperature
56	Temperature_Heater_16	Temperature
57	Temperature_Heater_17	Temperature
58	Temperature_Heater_18	Temperature
59	Temperature_Heater_19	Temperature
60	Temperature_Heater_20	Temperature
61	Flow_Helium_1	Gas flow
62	Flow_Helium_2	Gas flow
63	Flow_Helium_3	Gas flow
64	Flow_Helium_4	Gas flow
65	Flow_Helium_5	Gas flow
66	Flow_Helium_6	Gas flow
67	Flow_Helium_7	Gas flow
68	Logistic_2	Logistics
69	Logistic_3	Logistics
70	Flow_Clean_1	Gas flow
71	Flow_Clean_2	Gas flow
72	Flow_Clean_3	Gas flow
73	Flow_Clean_4	Gas flow
74	Flow_Clean_5	Gas flow
75	Flow_Clean_6	Gas flow
76	Flow_Nitrousoxide_1	Gas flow
77	Flow_Nitrousoxide_2	Gas flow
78	Flow_Nitrousoxide_3	Gas flow
79	Flow_Nitrousoxide_4	Gas flow
80	Flow_Nitrousoxide_5	Gas flow
81	Flow_Nitrousoxide_6	Gas flow
82	Flow_Nitrousoxide_7	Gas flow
83	Flow_Nitrousoxide_8	Gas flow
84	Flow_Nitrousoxide_9	Gas flow
85	Flow_Nitrousoxide_10	Gas flow
86	Flow_Nitrousoxide_11	Gas flow

Feature Overview (1) for AMAT Producer

Index	Feature	Category	Index	Feature	Category
87	Flow_Nitrousoxide_12	Gas flow	133	Counter_20	Counter
88	Flow_Nitrousoxide_13	Gas flow	134	Counter_21	Counter
89	Flow_Nitrousoxide_14	Gas flow	135	Time_1	Time
90	Flow_Nitrousoxide_15	Gas flow	136	Power_Bias_9	Power
91	Flow_Nitrousoxide_16	Gas flow	137	Power_Bias_10	Power
92	Flow_Nitrousoxide_17	Gas flow	138	Power_Bias_11	Power
93	Flow_Nitride_1	Gas flow	139	Power_Bias_12	Power
94	Flow_Nitride_2	Gas flow	140	Power_Bias_13	Power
95	Flow_Nitride_3	Gas flow	141	Power_Bias_14	Power
96	Flow_Nitride_4	Gas flow	142	Power_Bias_15	Power
97	Flow_Nitride_5	Gas flow	143	Power_Bias_16	Power
98	Flow_Nitride_6	Gas flow	144	Power_Bias_17	Power
99	Flow_Nitride_7	Gas flow	145	Power_Bias_18	Power
100	Flow_Nitride_8	Gas flow	146	Power_Bias_19	Power
101	Flow_Nitride_9	Gas flow	147	Power_Bias_20	Power
102	Flow_Nitride_10	Gas flow	148	Power_Bias_21	Power
103	Flow_Nitrogentriflouride_1	Gas flow	149	Power_Bias_22	Power
104	Flow_Nitrogentriflouride_2	Gas flow	150	Power_Bias_23	Power
105	Flow_Nitrogentriflouride_3	Gas flow	151	Power_Bias_24	Power
106	Flow_Nitrogentriflouride_4	Gas flow	152	Power_Bias_25	Power
107	Flow_Nitrogentriflouride_5	Gas flow	153	Power_Bias_26	Power
108	Flow_Nitrogentriflouride_6	Gas flow	154	Power_Bias_27	Power
109	Flow_Nitrogentriflouride_7	Gas flow	155	Power_Bias_28	Power
110	Flow_Ammonia_1	Gas flow	156	Power_Bias_29	Power
111	Flow_Ammonia_2	Gas flow	157	Power_Bias_30	Power
112	Flow_Ammonia_3	Gas flow	158	Power_Bias_31	Power
113	Flow_Ammonia_4	Gas flow	159	Power_Bias_32	Power
114	Flow_Ammonia_5	Gas flow	160	Power_Bias_33	Power
115	Counter_2	Counter	161	Flow_Silane_1	Gas flow
116	Counter_3	Counter	162	Flow_Silane_2	Gas flow
117	Counter_4	Counter	163	Flow_Silane_3	Gas flow
118	Counter_5	Counter	164	Flow_Silane_4	Gas flow
119	Counter_6	Counter	165	Flow_Silane_5	Gas flow
120	Counter_7	Counter	166	Flow_Silane_6	Gas flow
121	Counter_8	Counter	167	Flow_Silane_7	Gas flow
122	Counter_9	Counter	168	Flow_Silane_8	Gas flow
123	Counter_10	Counter	169	Flow_Silane_9	Gas flow
124	Counter_11	Counter	170	Flow_Silane_10	Gas flow
125	Counter_12	Counter	171	Flow_Silane_11	Gas flow
126	Counter_13	Counter	172	Flow_Silane_12	Gas flow
127	Counter_14	Counter	173	Flow_Silane_13	Gas flow
128	Counter_15	Counter	174	Flow_Silane_14	Gas flow
129	Counter_16	Counter	175	Pressure_Throttle_1	Pressure
130	Counter_17	Counter	176	Pressure_Throttle_2	Pressure
131	Counter_18	Counter	177	Pressure_Throttle_3	Pressure
132	Counter_19	Counter	178	Pressure_Throttle_4	Pressure

Feature Overview (2) for AMAT Producer

Index	Feature	Category
179	Pressure_Throttle_5	Pressure
180	Pressure_Throttle_6	Pressure
181	Pressure_Throttle_7	Pressure
182	Pressure_Throttle_8	Pressure
183	Time_2	Time
184	Time_3	Time
185	Time_4	Time
186	Time_5	Time
187	Time_6	Time
188	Time_7	Time
189	Time_8	Time
190	Time_9	Time
191	Time_10	Time
192	Time_11	Time
193	Time_12	Time
194	Time_13	Time
195	Time_14	Time
196	Time_15	Time
197	Time_16	Time
198	Time_17	Time
199	Artificial_1_ran	Artificial
200	Artificial_2_ran	Artificial
201	Artificial_3_ran	Artificial
202	Artificial_4_cor	Artificial
203	Artificial_5_cor	Artificial

Feature Overview (3) for AMAT Producer

A.5 ERBE Feature Selection Parameters

Param	Category	Value	Description
nRn	Artificial	2	Number of artificial features with normal distribution
nRu	Artificial	1	Number of artificial features with uniform distribution
nRc	Artificial	2	Number of highly correlated features
$splitTr$	Split ratio	0.8	Ratio of data used for training
$splitVal$	Split ratio	0.2	Ratio of data used for validation
$splitTst$	Split ratio	0	Ratio of data used for testing
ker	SVR	rbf	SVR kernel type
$loss$	SVR	eInsen	SVR loss function type: ε-insensitive
ε	SVR	0.01	SVR epsilon for e-insensitive loss
$rmse$	Evaluation	-	Root Mean Squared Error (RMSE)
mae	Evaluation	-	Mean Absolute Error (MAE)
$r2$	Evaluation	-	Coefficient of Determination (R^2)
$yMin$	Normalization	-	Minimum of target Y
$yMax$	Normalization	-	Maximum of target Y
n	Dataset	-	Number of instances within dataset
m	Dataset	-	Number of features within dataset
nG	GA	25	Number of generations during GA optimization
nI	GA	5	Number of populated individuals during ERBE GA
mR	GA	0.05	Ratio of original features to be flipped during GA cycle
rR	GA	0.1	Ratio of eliminated original features during each GA stage
$nAIF$	GA	80	Number of initial/original features
aF	GA	-	Currently active feature subset

Overview of all parameters used for ERBE Feature Selection

The entire computation of the new ERBE FS algorithm is performed using an Intel i5 2.6 GHz dual core processor with 4 GB memory and a 32-bit Windows 7 Enterprise OS installed.

A.6 ERBE Results for AMAT Producer

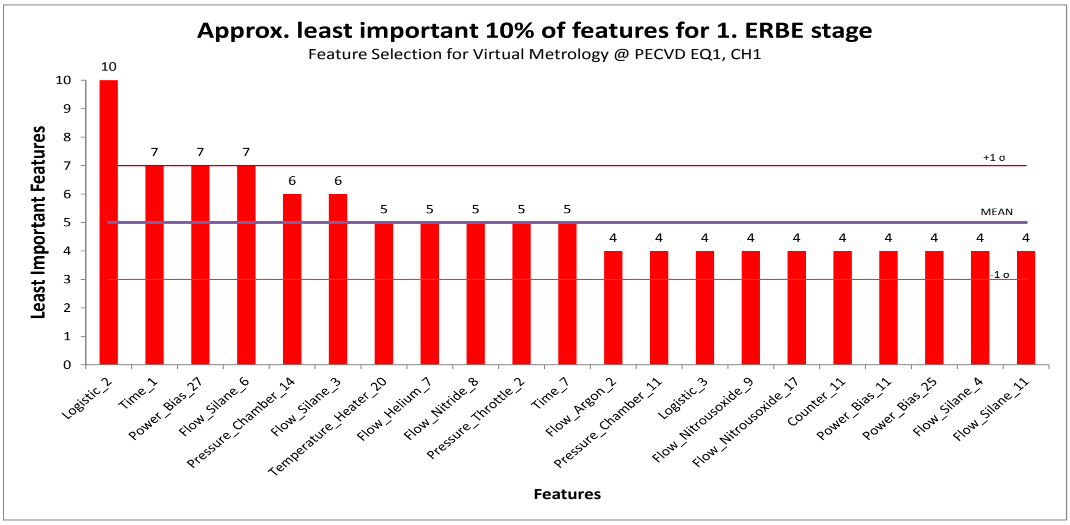


Figure A.1: ERBE stage 1 illustrating features selected within approximate least important 10 % and yielding a CV(RMSE) of 1.7577 is designed for fast elimination of features contributing mainly noise and inferring variables. Least important and removed real features (red) are selected by LOO FS.

ERBE Stage 1: The results in terms of the least important features for the first ERBE stage are displayed in figure A.1. None of the five artificial features are selected. Thus, the initial feature set contains many very noisy and inferring features. Features of various categories are revealed as dispensable while the range how often features are selected is spread from ten down to four. Furthermore, the standard deviation of two around the mean of five approves the high number of dispensable information in the 203 features. Finally, Logistic_2 is clearly recognized as most unimportant feature.

ERBE Stage 2: The results in terms of the least important features for the second ERBE stage are displayed in figure A.2. Already 4 out of the 5 artificial features are selected frequently whereat in fact it is interesting to recognize that the correlated complements Artificial_3_ran and Artificial_5_cor are selected both 8 times within the least important 10 % of the features. A feature set still containing many noisy features is most likely due to a small standard deviation of 1 around the mean of 7. Features of various categories are revealed as dispensable while the range how often features are selected is decreased to 8 down to 5. 3 process parameters (i. e. Flow_Nitrousoxide_10, Flow_Nitrogentriflouride_5 & Power_Bias_30) are observed as most

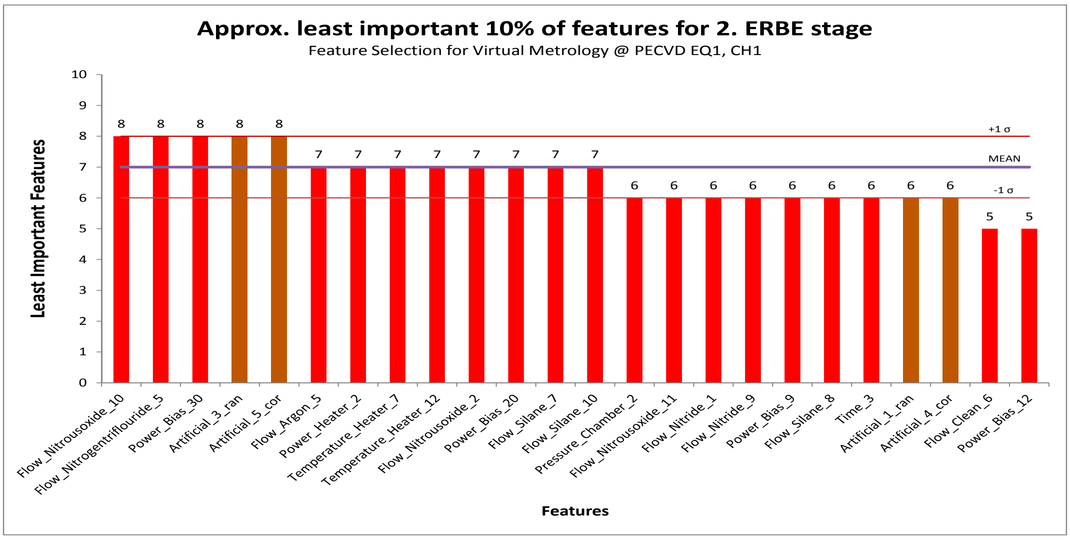


Figure A.2: ERBE stage 2 illustrating features selected within approximate least important 10 % and yielding a CV(RMSE) of 1.5554 is designed for fast elimination of features contributing mainly noise and inferring variables compared to artificial features (burnt orange). Least important and removed real features (red) are selected by LOO FS.

unimportant. Due to transition criterion 1.1 (0.73 = 2 artificial features 8 times selected + 2 x 6 / 3 real features 8 times selected + 2 x 7) no transition to ERBE part II is made.

ERBE Stage 3: The results in terms of the least important features for the third ERBE stage are displayed in figure A.3. Again, 4 out of the 5 artificial features are selected frequently. A feature set still containing many noisy features is most likely due to a small standard deviation of 1 around the mean of 6. Features of various categories are revealed as dispensable while the range how often features are selected is decreased again to 7 down to 5 indicating even less noise. 5 process parameters (i. e. Pressure_Chamber_3, Pressure_Chamber_6, Power_Heater_8, Flow_Nitride_2 & Time_8) are observed as most unimportant. Transition criteria 1.1 (0.8 = 3 artificial features 6 times selected + 2 x 5 / 5 real features 7 times selected) and 1.2 (5 out of 5 features) are both met and thus ERBE part II is executed subsequently.

ERBE Stage 4: The results in terms of the least important features for the fourth ERBE stage are displayed in figure A.4. Only 3 out of the 5 artificial features are selected. Almost all features are selected 18 times due to a very dense feature distribution yielding a mean of 18 and a standard deviation of 0. Only two real process parameters Power_Bias_14 and Flow_Silane_13 of different categories are revealed as less informative. Due to the fact that the artificial features are not grouped together at the beginning of ERBE part II criterion 2.2 is not considered for the transition to ERBE part III. Transition criterion 2.1 (0.6 = 2 artificial features 19 times

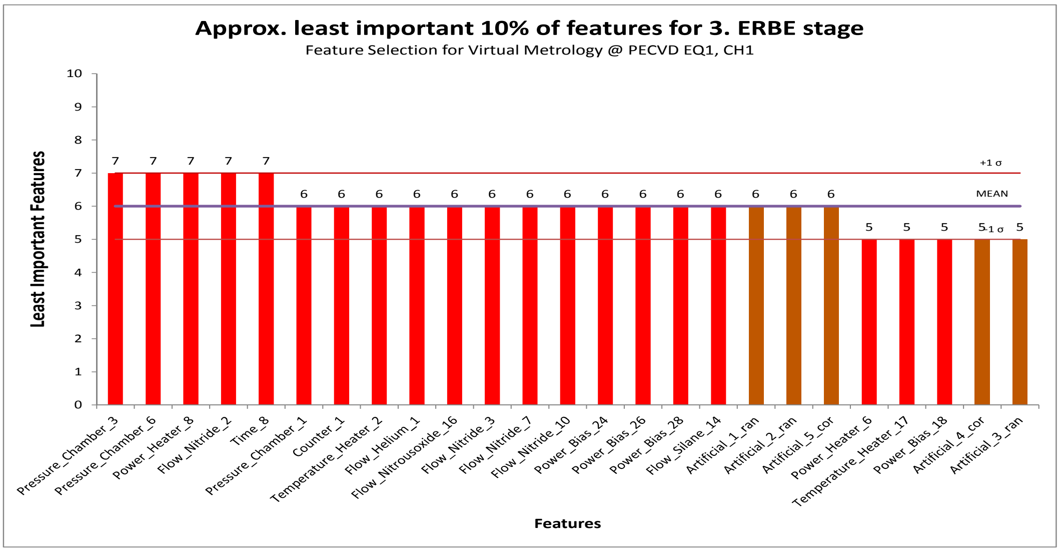


Figure A.3: ERBE stage 3 illustrating features selected within approximate least important 10 % and yielding a CV(RMSE) of 1.4961 is designed for fast elimination of features contributing mainly noise and inferring variables compared to artificial features (burnt orange). Least important and removed real features (red) are selected by LOO FS.

selected + 1 x 18 / 2 real features 19 times selected + 3 x 18) is not met and thus ERBE part II is still executed.

ERBE Stage 5: The results in terms of the least important features for the fifth ERBE stage are displayed in figure A.5. Again, only 3 out of the 5 artificial features are selected. Less obvious differentiation between most important and noisy process parameters in the feature subset is most likely with a small standard deviation of 1 around the mean of 18. Features of various categories are revealed as dispensable while almost all features are selected 18 or 19 times with one exception of artificial variable 1 selected 21 times. 9 real process parameters of different categories are revealed as most unimportant. Due to the fact that the artificial features are not grouped together at the beginning of ERBE part II criterion 2.2 is not considered for the transition to ERBE part III. Transition criterion 2.1 (0.6 = 1 artificial features 21 times selected + 2 x 18 / 5 real features 19 times selected) is not met and thus ERBE part II is still executed.

ERBE Stage 6: The results in terms of the least important features for the sixth ERBE stage are displayed in figure A.6. Slightly improved differentiation between more important and noisy process parameters is visible with a small standard deviation of 1 around the mean of 19. Pressure_Chamber_7 as only features is revealed as most unimportant while again most features are selected 18 or 19 times. Due to the fact that the artificial features are not grouped together at the beginning of ERBE part II criterion 2.2 is not considered for the transition to

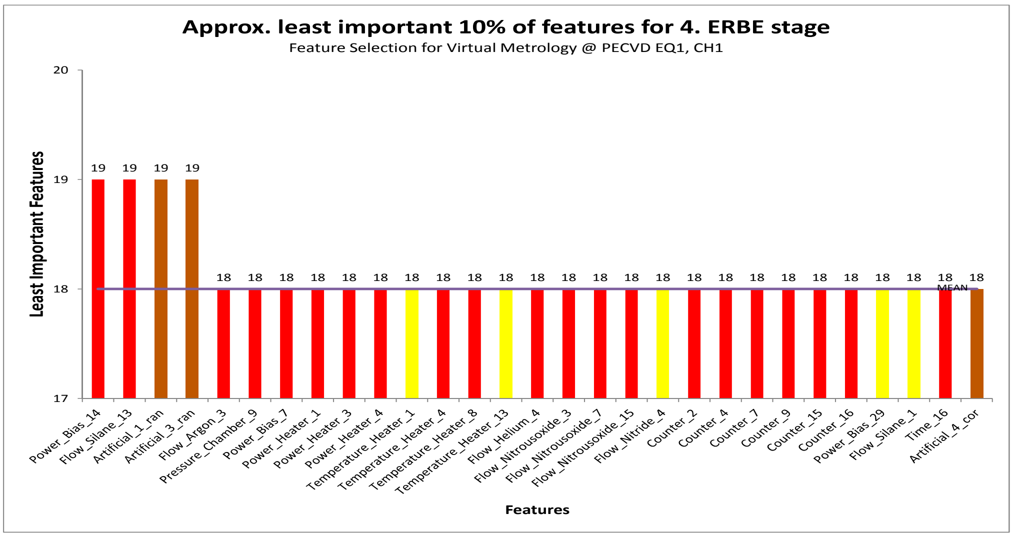


Figure A.4: ERBE stage 4 illustrating features selected within approximate least important 10 % and yielding a CV(RMSE) of 1.3507 is designed for feature subset optimization incorporating crucial interdependencies by GA FS. Least important and removed real features (red) are differentiated from others randomly surviving features (yellow) and artificial variables (burnt orange).

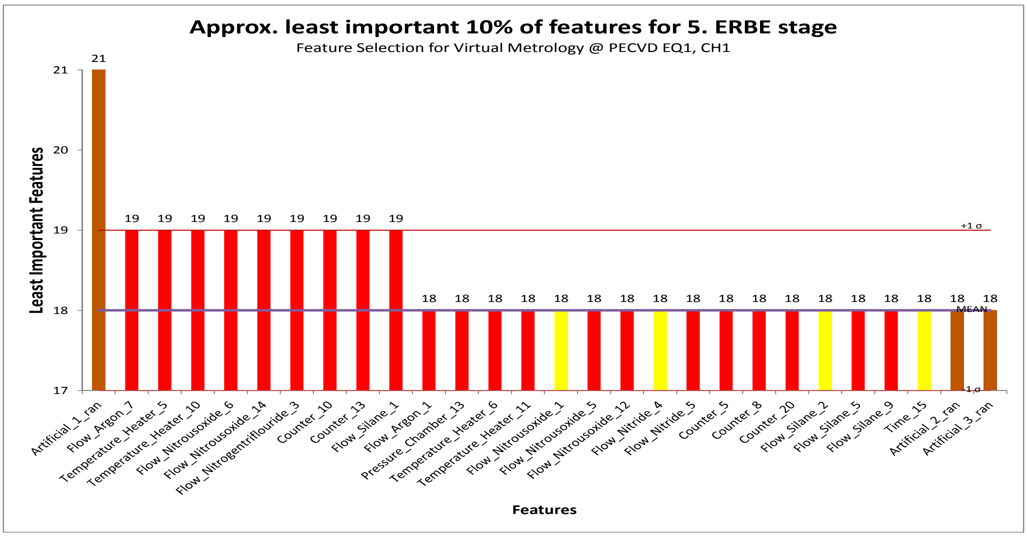


Figure A.5: ERBE stage 5 illustrating features selected within approximate least important 10 % and yielding a CV(RMSE) of 1.3297 is designed for feature subset optimization incorporating crucial interdependencies by GA FS. Least important and removed real features (red) are differentiated from others randomly surviving features (yellow) and artificial variables (burnt orange).

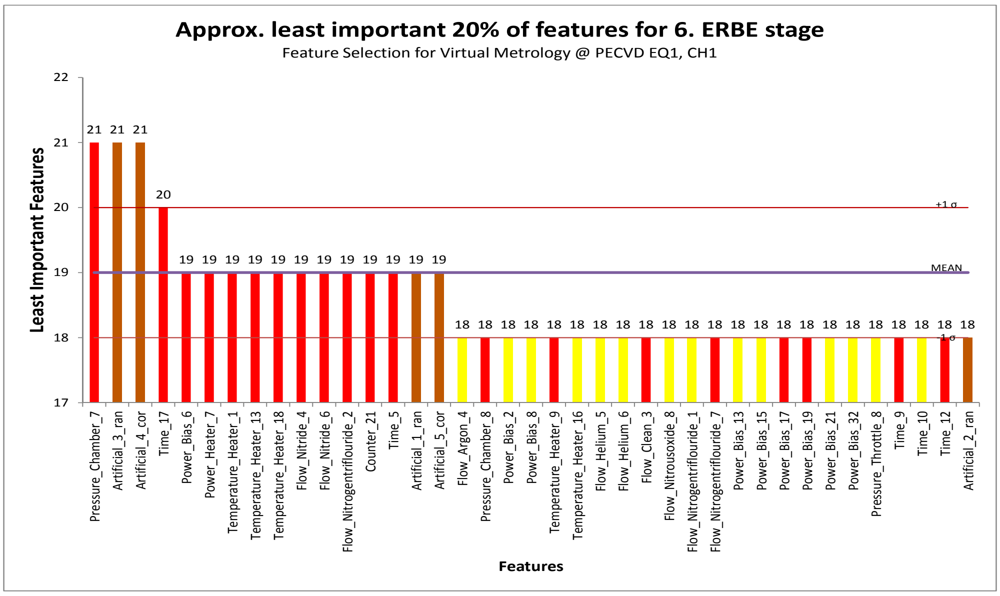


Figure A.6: ERBE stage 6 illustrating features selected within approximate least important 20 % and yielding a CV(RMSE) of 1.3175 is designed for feature subset optimization incorporating crucial interdependencies by GA FS. Least important and removed real features (red) are differentiated from others randomly surviving features (yellow) and artificial variables (burnt orange).

ERBE part III. Transition criterion 2.1 (1 = 2 artificial features 21 times selected + 2 x 19 + 1 x 18 / 1 real features 21 times selected + 1 x 20 + 3 x 19) is exactly met and thus ERBE part III is executed subsequently.

ERBE Stage 7: The results in terms of the least important features for the seventh ERBE stage are displayed in figure A.7. Good differentiation and fine tuning feature optimization by LOO FS between most important and noisy process parameters in the feature subset is observed with a small standard deviation of 1 around the mean of 8. Features of various categories are revealed as dispensable. 5 real process parameters of different categories (i. e. Power_Bias_5, Flow_Nitrogentriflouride_3, Flow_Nitrogentriflouride_6, Flow_Silane_2 & Pressure_Throttle_7) are revealed as most unimportant.

ERBE Stage 8: The results in terms of the least important features for the eighth ERBE stage are displayed in figure A.8. Slight differentiation and fine tuning feature optimization by LOO FS between most important and noisy process parameters in the feature subset is observed with a small standard deviation of 1 around the mean of 9. Features of various categories are revealed as dispensable. 6 real process parameters of different categories (i. e. Flow_Nitrousoxide_1, Counter_6, Counter_17, Counter_18, Power_Bias_13 & Flow_Silane_12) are revealed as most unimportant.

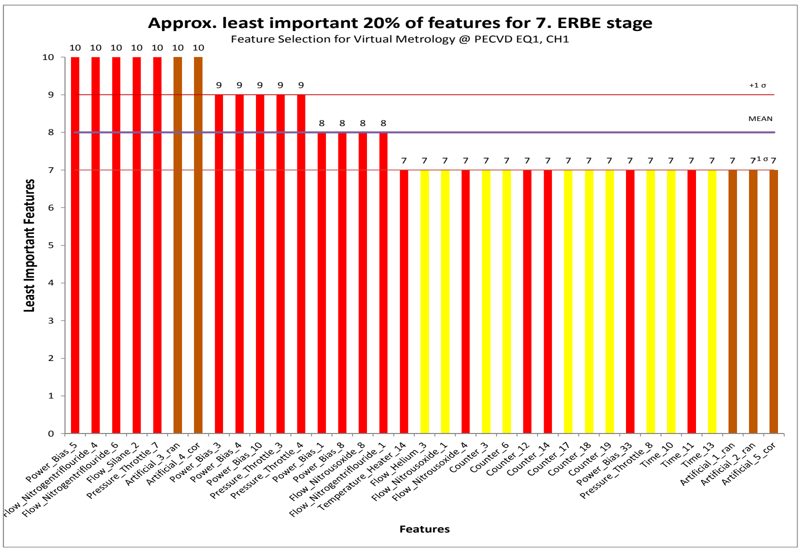


Figure A.7: ERBE stage 7 illustrating features selected within approximate least important 20 % and yielding a CV(RMSE) of 1.31 is designed for fine tuning feature optimization by LOO FS. Least important and removed real features (red) are differentiated from others randomly surviving features (yellow) and artificial variables (burnt orange).

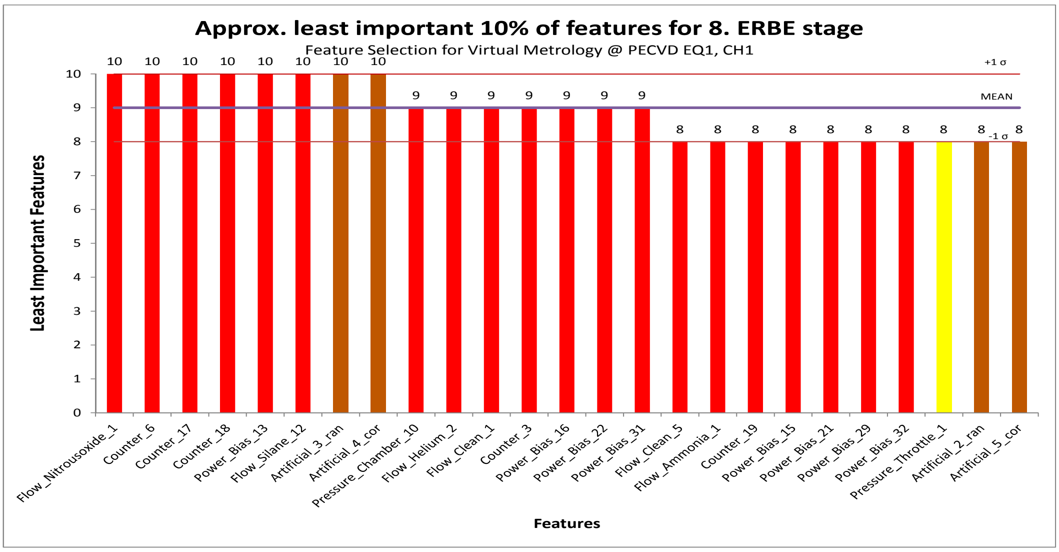


Figure A.8: ERBE stage 8 illustrating features selected within approximate least important 10 % and yielding a CV(RMSE) of 1.3065 is designed for fine tuning feature optimization by LOO FS. Least important and removed real features (red) are differentiated from others randomly surviving features (yellow) and artificial variables (burnt orange).

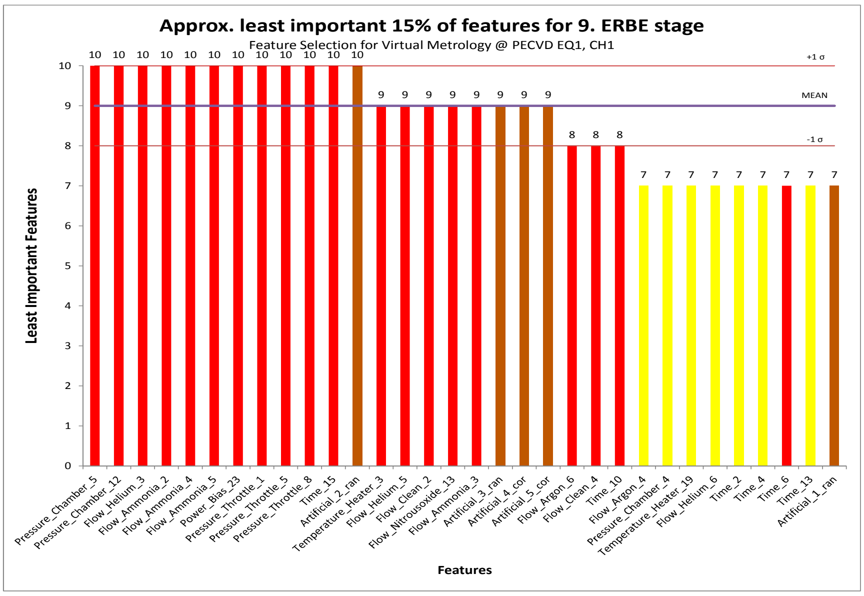


Figure A.9: ERBE stage 9 illustrating features selected within approximate least important 15 % and yielding a CV(RMSE) of 1.2649 is designed for fine tuning feature optimization by LOO FS. Least important and removed real features (red) are differentiated from others randomly surviving features (yellow) and artificial variables (burnt orange).

ERBE Stage 9: The results in terms of the least important features for the ninth ERBE stage are displayed in figure A.9. Slight differentiation and fine tuning feature optimization by LOO FS between most important and noisy process parameters in the feature subset is observed with a small standard deviation of 1 around the mean of 9. Features of various categories are revealed as dispensable. 11 real process parameters of different categories are revealed as most unimportant.

Glossary

AMAT	Applied Materials
APC	Advanced Process Control
BPNN	Back Propagation Neural Network
CBA	Cost-Benefit Analysis
CM	Configuration Module
CRISP-DM	CRoss Industrial Standard Process for Data Mining
CV(RMSE)	Coefficient of Variation of the RMSE
CVD	Chemical Vapor Deposition
DB	Database
DC	Direct Current
DM	Data Mining
DP	Data Preparation
DR	Deposition Rate
DS	Dataset
DT	Deposition Time
DTree	Decision Tree
ERBE	Evolutionary Repetitive Backward Elimination
ES	Expert Selection
fab	Fabrication Plant
FDC	Fault Detection and Classification
FS	Feature Selection
GA	Genetic Algorithm

HDP	High Density Plasma
IMPROVE	Implementing Manufacturing science solutions to increase equiPment pROductiVity and fab pErformance
LCL	Lower Control Limit
LOO	Leave-One-Out
LT	Layer Thickness
M5’	Decision Tree M5’
MAE	Mean Absolute Error
ML	Machine Learning
MLR	Multiple Linear Regression
MSE	Mean Squared Error
MW	Moving Window
NN	Neural Network
PCA	Principle Component Analysis
PECVD	Plasma Enhanced Chemical Vapor Deposition
PLS	Partial Least Squares
PTM	Prediction and Training Module
R^2	Coefficient of Determination
R2R	Run-To-Run
RBF	Radial Basis Function
RELIEF	FS Algorithm
RF	Radio Frequency
RFE	Recursive Feature Elimination
RI	Reliance Index
RMSE	Root Mean Squared Error
SM	Semiconductor Manufacturing
SS	Stepwise Selection

SVM	Support Vector Machine
SVR	Support Vector Regression
UCL	Upper Control Limit
VM	Virtual Metrology

Bibliography

[1] Almuallim, H. ; Dietterich, T. G.: Learning with many irrelevant features. In: *Proceedings of AAAI*, 1991, pp. 547–552

[2] Anandan, P. ; Varma, M. ; Joy, J.: *Multiple Kernel Learning*. http://research.microsoft.com/en-us/groups/vgv/vgv4.png. Version: 2012

[3] AppliedMaterials, Inc.: *Applied Material Technical Training: Ultima HDP-CVD Centura Process Optimization*, 1998. – The documentation is confidential and dedicated to customer's internal use only, therefore not publicly accessible

[4] AppliedMaterials, Inc.: *Ultima HDP-CVD Centura Process: Optimization & Troubleshooting*, 1999. – The documentation is confidential and dedicated to customer's internal use only, therefore not publicly accessible

[5] AppliedMaterials, Inc.: *Producer 200 mm/300 mm Functional Description for PE TEOS, PE Silane, SACVD and Black Diamond*, 2000. – The documentation is confidential and dedicated to customer's internal use only, therefore not publicly accessible

[6] Baek, K. ; Song, K. ; C., Han. ; Choi, G. ; Cho, H. ; Edgar, T.: Implementation of a robust virtual metrology for plasma etching through effective variable selection and recursive update technology. In: *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures* 32 (2014), Nr. 1, pp. 1–10

[7] Bai, Q.: Analysis of Particle Swarm Optimization Algorithm. In: *Computer and information science* 3 (2010), Nr. 1, pp. 180

[8] Barzilay, O. ; Brailovsky, V. L.: On Domain Knowledge and Feature Selection using a Support Vector Machine. In: *Pattern Recognition Letters* 20 (1999), Nr. 5, pp. 475–484

[9] Bennett, K. P. ; Mangasarian, O. L.: Robust Linear Programming Discrimination of two Linearly Inseparable Sets. In: *Optimization Methods and Software* 1 (1992), Nr. 1, pp. 23–34

[10] Bi, J. ; Bennett, K. P. ; Embrechts, M. ; Breneman, C. ; Song, M.: Dimensionality reduction via sparse support vector machines. In: *Journal of Machine Learning Research* 3 (2003), pp. 1229–1243

[11] Bishop, C. M. et al.: *Pattern Recognition and Machine Learning*. Bd. 1. Springer New York, 2006

[12] Blum, A. L. ; Langley, P.: Selection of Relevant Features and Examples in Machine Learning. In: *Artificial Intelligence* 97 (1997), pp. 245–271

[13] Braun, M.: *Temperaturprozessdatenerfassung mittels Infrarotsensorik bei der HDP-CVD Abscheidung*, University of Applied Science Regensburg, Bachelor Thesis, 2011

[14] Cameron, A. C. ; Windmeijer, F. A. G.: An R-squared measure of goodness of fit for some common nonlinear regression models. In: *Journal of Econometrics* 77 (1997), Nr. 2, pp. 329–342

[15] Cardie, C.: Using decision tree to improve case-based learning. In: *Proceedings of the 10th International Conference on Machine Learning*, 1993, pp. 25–32

[16] Carriquiry, A.: *Regression: Inference - part 3.* Lecture, 2004. – Iowa State University

[17] Cassotti, M. ; Grisoni, F.: *Variable Selection Methods: An Introduction*, 2007. – Milano Chemometrics and QSAR Research Group

[18] Chang, J. Y.-C. ; Cheng, F.-T.: Application Development of Virtual Metrology in Semiconductor Industry. In: *IECON 31st Annual Conference of IEEE Industrial Electronics Society*, 2005, pp. 124–129

[19] Chang, Y.-J. ; Kang, Y. ; Hsu, C.-L. ; Chang, C.-T. ; Chan, T.-Y.: Virtual Metrology Technique for Semiconductor Manufacturing. In: *2006 International Joint Conference on Neural Networks*, 2006, pp. 5289–5293

[20] Chapman, P. ; Clinton, J. ; Kerber, R. ; Khabaza, T. ; Reinartz, T. ; Shearer, C. ; Wirth, R.: *CRISP-DM 1.0 Step-by-step data mining guide.* CRISP-DM consortium, 2000

[21] Chen, P. H. ; Wu, S. ; Lin, J. ; Ko, F. ; Wang, J. ; Yu, C. H. ; Liang, M. S.: Virtual Metrology: A solution for wafer to wafer advanced process control. In: *Proceedings of IEEE ISSM*, 2005, pp. 155–157

[22] Cheng, F.-T. ; Chang, J. Y.-C. ; Huang, H.-C. ; Kao, C.-A. ; Chen, Y.-L. ; Peng, J.-L.: Benefit Model of Virtual Metrology and Integrating AVM into MES. In: *Transactions on Semiconductor Manufacturing* 24 (2011), May, Nr. 2, pp. 261–272

[23] Cheng, F.-T. ; Chang, Y.-C. J.: Configuring AVM as a MES Component. In: *Proceedings of the 2010 IEEE/SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 2010, pp. 226–231

[24] Cheng, F.-T. ; Chen, Y.-T. ; Su, Y.-C. ; Zeng, D.-L.: Method for Evaluating Reliance Level of a Virtual Metrology System. In: *2007 IEEE International Conference on Robotics and Automation*, 2007, pp. 1590–1596

[25] Cheng, F.-T. ; Huang, H.-C. ; Wu, W.-M.: Dual-Phase Virtual Metrology Scheme. In: *Transactions on Semiconductor Manufacturing* 20 (2007), November, Nr. 4, pp. 566–571

[26] Colliez, J. ; Dufrenois, F. ; Hamad, D.: Robust Regression and Outlier Detection with SVR: Application to Optic Flow Estimation. In: *BMVC*, 2006, pp. 1229–1238

[27] Cortes, C. ; Vapnik, V:: Support-Vector Networks. In: *Machine learning* 20 (1995), Nr. 3, pp. 273–297

[28] CRISP-DM-consortium: *CRoss Industrial Standard Process for Data Mining.* http://www.crisp-dm.org

[29] Cristianini, N. ; Shawe-Taylor, J.: An Introduction to Support Vector Machines. (2000)

[30] Daelemans, W. ; Gillis, S. ; Durieux, G.: The Acquisition of Stress: a data-oriented Approach. In: *Computational Linguistics* 20 (1994), Nr. 3, pp. 421–451

[31] Daelemans, W. ; Hoste, V. ; De Meulder, F. ; Naudts, B.: Combined Optimization of feature selection and algorithm parameter interaction in machine learning of language. In: *Proceedings of the 14th European Conference on Machine Learning (ECML-2003)*, 2003

[32] Dorigo, M.: Ant Colony Optimization. In: *IEEE Computational Intelligence Magazine* 1 (2006), Nr. 4, pp. 28–39

[33] Duda, R. O. ; Hart, P. E. ; Stork, D. G.: *Pattern Recognition.* New York : Wiley, 2001

[34] Eads, D. ; Hill, D. ; Davis, S. ; Perkins, S. ; Ma, J. ; Porter, R. ; Theiler, J.: Genetic Algorithms and Support Vector Machines for Time Series Classification. In: *Proceedings of SPIE* Bd. 4787, 2002

[35] Eglese, R. W.: Simulated annealing: a tool for operational research. In: *European journal of operational research* 46 (1990), Nr. 3, pp. 271–281

[36] Eisele, I.: *Grundlagen der Silizium-Halbleitertechnologie.* Universität der Bundeswehr München, 2010

[37] Fahrmeir, L. ; Kneib, T. ; Lang, S.: *Regression: Modelle, Methoden und Anwendungen.* Berlin, Heidelberg : Springer-Verlag, 2009

[38] Fayyad, U. ; Piatetsky-Shapiro, G. ; Smyth, P.: From Data Mining to Knowledge Discovery in Databases. In: *AI Magazine* 17 (1996), Nr. 3, pp. 37–54

[39] Ferreira, A. ; Roussy, A. ; Conde, L.: Virtual Metrology Models for Predicting Physical Measurement in Semiconductor Manufacturing. In: *IEEE/SEMI Advanced Semiconductor Manufacturing Conference 2009*, 2009, pp. 149–154

[40] Ferreira, A. ; Roussy, A. ; Kernaflen, C. ; Gleispach, D. ; Hayderer, G. ; Gris, H. ; Besnard, J.: Virtual Metrology Models for Predicting Average PECVD Oxide Film

Thickness. In: *IEEE/SEMI Advanced Semiconductor Manufacturing Conference 2011*, 2011, pp. 1–6

[41] Fink, E.: *Computer-science quotes.* http://www.cs.cmu.edu/~eugene/quotes/prog.html

[42] Foerg, R.: Mikrosystemtechnik - Grundlagen integrierter Bauelemente / Hochschule Deggendorf. 2009 (1-4). – Lecture

[43] Froehlich, H. ; Chapelle, O. ; Schölkopf, B.: Feature Selection for Support Vector Machines by Means of Genetic Algorithms. In: *International Journal on Artificial Intelligence Tools*, IEEE Computer Society, 2003, pp. 142–148

[44] Gong, B. ; Guo, Z. ; Li, J. ; Zhu, G. ; Lv, S. ; Rao, S. ; Li, X.: *Lecture Notes in Computer Science.* Bd. 3614: *Application of a Genetic Algorithm - Support Vector Machine Hybrid for Prediction of Clinical Phenotypes Based on Genome-Wide SNP Profiles of Sib Pairs.* 1st Edition. Springer Berlin Heidelberg, 2005

[45] Grill, A.: Plasma-deposited diamondlike carbon and related materials. In: *IBM Journal of Research and Development* 43 (1999), Nr. 1-2, pp. 147–161

[46] Gu, C.: *Smoothing Spline ANOVA Models.* Springer, 2013

[47] Gunes, T. ; Polat, E.: Feature Selection for Multi-SVM Classifiers in Facial Expression Classification. In: *International Symposium on Computer and Information Sciences*, 2008

[48] Guyon, I. ; Elisseeff, A.: An introduction to variable and feature selection. In: *The Journal of Machine Learning Research* 3 (2003), pp. 1157–1182

[49] Guyon, I. ; Weston, J. ; Barnhill, S. ; Vapnik, V.: Gene Selection for Cancer Classification using Support Vector Machines. In: *The Journal of Machine Learning* 46 (2002), pp. 389–422

[50] Hall, M. A.: *Correlation-based Feature Selection for Machine Learning*, The University of Waikato, Hamilton, NewZealand, Diss., 1999

[51] Harrell, Frank E.: *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis.* Springer, 2001

[52] Hastie, T. ; Tibshirani, R. ; Friedman, J.: *The Elements of Statistical Learning - Data Mining, Inference, and Prediction.* 2nd Edition. Springer, 2009

[53] Hoeckele, Uwe: *Training Documents of the CVD Team of Infineon Technologies AG*, 1998. – The documentation is confidential and dedicated to customer's internal use only, therefore not publicly accessible

[54] Hofmann, H. ; Spindler, J.: *Verfahren in der Beschichtungs- und Oberflächentechnik: Grundlagen - Vorbehandlung - Oberflächenreaktionen - Schichtabscheidung - Strukturierung - Prüfung.* 2nd Edition. Hanser, 2010

[55] Hollauer, C.: *Modelling of Thermal Oxidation and Stress Effects*, The Vienna University, Diss., 2007

[56] Huang, C.-L. ; Wang, C.-J.: A GA-based Feature Selection and Parameters Optimization for Support Vector Machines. In: *Expert Systems with Applications* 31 (2006), pp. 231–240

[57] Huang, H.-C. ; Su, Y.-C. ; Cheng, F.-T. ; Jian, J.-M.: Development of a Generic Virtual Metrology Framework. In: *Proceedings of the IEEE Conference on Automation Science and Engineering*, 2007, pp. 282–287

[58] Huang, Y.-T. ; Cheng, F.-T. ; Chen, Y.-T.: Importance of Data Quality in Virtual Metrology. In: *IEEE Industrial Electronics, IECON 2006-32nd Annual Conference on*, 2006, pp. 3727–3732

[59] Huang, Y.-T. ; Huang, H.-C. ; Cheng, F.-T. ; Liao, T.-S. ; Chang, F.-C.: Automatic Virtual Metrology System Design and Implementation. In: *4th IEEE Conference on Automation Science and Engineering*, 2008, pp. 223–229

[60] Hung, M.-H. ; Huang, H.-C. ; Yang, H.-C. ; Cheng, F.-T: Development of an Automatic Virtual Metrology Framework for TFT-LCD Industry. In: *6th IEEE Conference on Automation Science and Engineering*, 2010, pp. 879–884

[61] Hwang, S. ; Jeong, M. K. ; Yum, B.-J.: Robust Relevance Vector Machine With Variational Inference for Improving Virtual Metrology Accuracy. In: *IEEE Transaction on Semiconductor Manufacturing* 27 (2014), pp. 83–94

[62] Imai, S. ; Sato, N. ; Kitabata, M. ; Yasuda, S.: Fab-wide Equipment Monitoring and FDC System. In: *ISSM 2006 IEEE International Symposium on Semiconductor Manufacturing*, 2006, pp. 114–117

[63] IMPROVE-consortium: *Implementing Manufacturing science solutions to increase equiPment pROductiVity and fab pErformance.* http://www.eniac-improve.eu/

[64] Infineon, Technologies: *Infineon Company Presentation.* http://www.infineon.com/dgdl/IFX_2013_Q1_de_web.pdf?folderId=db3a304314dca38901154706908d19ba&fileId=db3a30432fbc32ee012fbf7aace53a74. Version: 2013

[65] Infineon, Technologies: *Infineon Company.* http://www.infineon.com/cms/en/corporate/company/index.html. Version: 2014

[66] Infineon Technologies AG, München: *Halbleiter: Technische Erläuterungen, Technologien und Kenndaten.* 3., überarbeitete und wesentlich erweiterte Auflage. Erlangen : Publicis Corporate Publishing, 2004

[67] ITRS: *International Technology Roadmap for Semiconductors: Factory Integration.* 2011 Edition. 2011

[68] John, G. H. ; Kohavi, R. ; Pfleger, K.: Irrelevant Feature and the Subset Selection Problem. In: *Proceedings of Eleventh International Conference on Machine Learning*, 1994, pp. 121–129

[69] Kang, P. ; Kim, D. ; Lee, H.-J. ; Doh, S. ; Cho, S.: Virtual Metrology for Run-to-Run Control in Semiconductor Manufacturing. In: *Expert Systems with Applications* 38 (2011), Nr. 3, pp. 2508–2522

[70] Kang, P. ; Lee, H.-J. ; Cho, S. ; Kim, D. ; Park, J. ; Park, C.-K. ; Doh, S.: A Virtual Metrology System for Semiconductor Manufacturing. In: *Expert Systems with Applications* 36 (2009), Nr. 10, pp. 12554–12561

[71] Kao, C.-A. ; Cheng, F.-T. ; Wu, W.-M.: Preliminary Study of Run-to-Run Control Utilizing Virtual Metrology with Reliance Index. In: *2011 IEEE International Conference on Automation Science and Engineering*, 2011, pp. 256–261

[72] Karagiannopoulos, M. ; Anyfantis, D. ; Kotsiantis, S. B. ; Pintelas, P. E.: Feature Selection for Regression Problems. In: *Proceedings of Hellenic European Research on Computer Mathematics & its Applications Conference (HERCMA)*, 2007

[73] Kennedy, J.: Particle Swarm Optimization. In: *Encyclopedia of Machine Learning*. Springer, 2010, pp. 760–766

[74] Khan, A. A. ; Moyne, J. R. ; Tilbury, D. M.: An Approach for factory-wide control utilizing virtual metrology. In: *Transactions on Semiconductor Manufacturing* 20 (2007), November, Nr. 4, pp. 364–375

[75] Khan, A. A. ; Moyne, J. R. ; Tilbury, D. M.: Virtual metrology and feedback control for semiconductor manufacturing processes using recursive partial least squares. In: *Journal of Process Control* 18 (2008), November, Nr. 10, pp. 961–974

[76] Kim, D. ; Kang, P. ; Cho, S. ; Lee, H.-J. ; Doh, S.: Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing. In: *Expert Systems with Applications* 39 (2011), Nr. 4, pp. 4075–4083

[77] Kim, S.: Margin-Maximized Redundancy-Minimized SVM-RFE for Diagnostic Classification of Mammograms. In: *Proceedings of the International Conference on Bioinformatics and Biomedicine Workshops*, 2011, pp. 562–569

[78] Kira, K. ; Rendell, L. A.: A practical approach to feature selection. In: *Proceedings of the Ninth International Conference on Machine Learning*, 1992, pp. 249–256

[79] Klemmt, A.: *Ablaufplanung in der Halbleiter- und Elektronikproduktion*. Wiesbaden : Springer Vieweg, 2012

[80] Koenig, A. ; Gratz, A.: *Advanced Techniques in Knowledge discovery and Data Mining*. Springer, 2005

[81] Kohavi, R. ; John, G. H.: Wrappers for Feature Subset Selection. In: *Artificial Intelligence* 97 (1997), Nr. 1, pp. 273–324

[82] Kohavi, R. ; Langley, P. ; Yun, Y.: The Utility of Feature Weighting in nearest-neighbor Algorithms. In: *Proceedings of the 9th European Conference on Machine Learning*, Springer, 1997, pp. 192–197

[83] Kohavi, R. ; Sommerfield, D.: Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology. In: *Proceedings 1st International Conference on Knowledge Discovery and Data Mining*, 1995, pp. 192–197

[84] Koitzsch, M. ; Merhof, J. ; Michl, M. ; Noll, H. ; Nemecek, A. ; Honold, A. ; Kleineidam, G. ; Lebrecht, H.: Implementing Virtual Metrology into Semiconductor Production Processes - an Investment Assessment. In: *Proceedings of the 2011 Winter Simulation Conference*, 2011, pp. 2017–2028

[85] Kubat, M. ; Flotzinger, D. ; Pfurtscheller, G.: Discovering Patterns in EEG Signals: Comparative Study of a few Methods. In: *Proceedings of the 1993 European Conference on Machine Learning*, 1993, pp. 367–371

[86] Langley, P. ; Sage, S.: Induction of Selective Bayesian Classifiers. In: *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, 1994, pp. 399–406

[87] Leicht, Carolin: *Analyse und Optimierung von Algorithmen des Maschinellen Lernens in der Virtuellen Messtechnik*, University Leipzig, Diploma Thesis, 2013

[88] Lenz, B.: *Smart Feature Selection to enable Advanced Virtual Metrology*, Eberhard Karls University at Tübingen, Germany, Diss., 2015

[89] Lenz, B. ; Barak, B.: *Generic Data Mining System in Semiconductor Manufacturing.* Poster, 2012. – 2012 IEEE 12th International Conference on Data Mining (ICDM)

[90] Lenz, B. ; Barak, B.: Data Mining and Support Vector Regression Machine Learning in Semiconductor Manufacturing to Improve Virtual Metrology. In: *2013 46th Hawaii International Conference on System Sciences (HICSS)* IEEE, 2013, pp. 3447–3456

[91] Lenz, B. ; Barak, B. ; Kyek, A.: *Integration of the IMPROVE framework at Infineon Technologies AG.* Poster, 2012. – 2012 Advanced Process Control and Manufacturing Conference (APCM)

[92] Lenz, B. ; Barak, B. ; Kyek, A. ; Purwins, H.: *Data Mining and Machine Learning Technique in Semiconductor Manufacturing Processes (PECVD).* Presentation, 2011. – International SEMATECH Manufacturing Initiative (ISMI)

[93] Lenz, B. ; Barak, B. ; Leicht, C. ; Muehrwald, J.: Virtual Metrology in Semiconductor Manufacturing by means of Predictive Machine Learning Models. In: *Proceedings of*

the 12th International Conference on Machine Learning and Application (ICMLA), 2013, pp. 174–177

[94] Lenz, B. ; Barak, B. ; Leicht, C. ; Muehrwald, J.: Development of Smart Feature Selection for advanced Virtual Metrology. In: *2014 25th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)* IEEE, 2014, pp. 145–150

[95] Li, F. ; Mi, H. ; Yang, F.: Exploring the Stability of Feature Selection for Imbalanced Intrusion Detection Data. In: *IEEE Proceedings of the Ninth International Conference on Control and Automation*, 2011

[96] Li, G.-Z. ; Meng, H.-H. ; Yang, M. Q. ; Yang, J. Y.: Combining Support Vector Regression with Feature Selection for Multivariate Calibration. In: *Neural Computing and Applications* 18 (2009), Nr. 7, pp. 813–820

[97] Li, L. ; Duan, Y.: A GA-based Feature Selection and Parameters Optimization for Support Vector Regression. In: *Proceedings of the Seventh International Conference on Natural Computation*, 2011, pp. 335–339

[98] Li, W. ; Zhao, Y. ; Song, Y. ; Yang, Z.: SVM Feature Selection and Sample Regression for Chinese Medicine Research. In: *Proceedings of the 2008 IEEE International Conference on Information and Automation*, 2008, pp. 1773–1777

[99] Li, X. ; Peng, S. ; Zhan, X. ; Zhang, J. ; Xu, Y.: Comparison of feature selection methods for multiclass cancer classification based on microarray data. In: *Proceedings of the International Conference on Biomedical Engineering and Informatics*, 2011, pp. 1692–1696

[100] Liebermann, M. A. ; Lichtenberg, A. J.: *Principles of Plasma Discharges and Materials Processing*. Wiley, 2005

[101] Lin, T.-H. ; Cheng, F.-T. ; Ye, A.-J. ; Wu, W.-M. ; Hung, M.-H.: A Novel Key-variable Sifting Algorithm for Virtual Metrology. In: *Proceedings of the International Conference on Robotics and Automation*, 2008, pp. 3636–3641

[102] Lin, T.-H. ; Hung, M.-H. ; Lin, R.-C. ; Cheng, F.-T.: A Virtual Metrology Scheme for Predicting CVD Thickness in Semiconductor Manufacturing. In: *Proceedings of the International Conference on Robotics and Automation*, 2006, pp. 1054–1059

[103] Liu, H. ; Motoda, H.: *Feature selection for knowledge discovery and data mining.* Springer, 1998

[104] Liu, H. ; Motoda, H. ; Yu, L.: Feature Selection with Selective Sampling. In: *In Proceedings of the Nineteenth International Conference on Machine Learning*, 2002, pp. 395–402

[105] Lynn, S. ; Ringwood, J. ; McGearailt, N.: Gaussian Process Regression for Virtual Metrology of Plasma Etch. In: *Proceedings of the 21h Irish Signals and Systems Conference (ISSC)*, 2010

[106] Lynn, S. ; Ringwood, J. ; McGearailt, N.: Global and Local Virtual Metrology Models for a Plasma Etch Process. In: *IEEE Transactions on Semiconductor Manufacturing* 25 (2012), Nr. 3

[107] Lynn, S. ; Ringwood, J. ; Ragnoli, E. ; McLoone, S. ; McGearailt, N.: Virtual Metrology for Plasma Etch using Tool Variables. In: *Proceedings of the 2009 IEEE/SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 2009, pp. 143–148

[108] Maldonado, S. ; Weber, R.: Feature Selection for Support Vector Regression via Kernel Penalization. In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 2010, pp. 1–7

[109] Manikas, Theodore W. ; Cain, James T.: Genetic algorithms vs. simulated annealing: A comparison of approaches for solving the circuit partitioning problem. (1996)

[110] Mao, Q. ; Tsang, I. W.-H.: Optimizing Performance Measures for Feature Selection. In: *11th IEEE International Conference on Data Mining*, 2011, pp. 1170–1175

[111] Mattes, A. ; Schöpka, U. ; Schellenberger, M. ; Scheibelhofer, P. ; Leditzky, G.: Virtual Equipment for benchmarking Predictive Maintenance Algorithms. In: *Proceedings of the 2012 Winter Simulation Conference*, 2012, pp. 1–12

[112] May, G. S. ; Spanos, C. J.: *Fundamentals of Semiconductor Manufacturing and Process Control*. Wiley. com, 2006

[113] Meidan, Y. ; Lerner, B. ; Rabinowitz, G. ; Hassoun, M.: Cycle-Time Key Factor Identification and Prediction in Semiconductor Manufacturing Using Machine Learning and Data Mining. In: *IEEE Transactions on Semiconductor Manufacturing* 24 (2011), Nr. 2

[114] Mendenhall, W. ; Sincich, T.: *A Second Course in Statistics: Regression Analysis*. 6th Edition. Pearson, 2003

[115] Miller, A.: *Subset Selection in Regression*. 2nd Edition. 2002

[116] Miller, M. T. ; Jerebko, A. K. ; Malley, J. D. ; Summers, R. M.: Feature Selection for Computer-Aided Polyp Detection using Genetic Algorithms. In: *Medical Imaging 2003*, 2003, pp. 102–110

[117] Mitchell, M.: An Introduction to Genetic Algorithms. In: *Cambridge, Massachusetts London, England, Fifth printing* 3 (1999)

[118] Moyne, J. ; Del Castillo, E. ; Hurwitz, A. M.: *Run-to-Run Control in Semiconductor Manufacturing*. CRC Press, 2010

[119] Moyne, J. R.: Making the move to fab-wide APC. In: *Solid State Technology* 47 (2004), September, Nr. 9, pp. 47–52

[120] Moyne, J. R.: A Blueprint for enterprise-wide Deployment of Advanced Process Control. In: *Solid State Technology* 52 (2009), September, Nr. 7

[121] Narendra, P. ; Fukunaga, K.: A Branch and Bound Algorithm for Feature Subset Selection. In: *IEEE Transactions on Computers* (1977)

[122] Pampuri, S. ; Schirru, A. ; Fazio, G. ; De Nicolao, G.: Multilevel Lasso applied to Virtual Metrology in Semiconductor Manufacturing. In: *IEEE Conference on Automation Science and Engineering (CASE)*, 2011

[123] Pampuri, S. ; Schirru, A. ; Susto, G. A. ; De Luca, C. ; Beghi, A. ; De Nicolao, G.: Multistep Virtual Metrology Approaches for Semiconductor Manufacturing Processes. In: *IEEE Conference on Automation Science and Engineering (CASE)*, 2012, pp. 91–96

[124] Platt, J. C.: Fast Training of Support Vector Machines using Sequential Minimal Optimization. (1999)

[125] Prakash, P. ; McLoone, S. F.: Plasma Etch Process Virtual Metrology using Aggregative Linear Regression. In: *2011 International Conference on Soft Computing and Pattern Recognition*, 2011, pp. 538–543

[126] Prakash, P. ; Schirru, A. ; Hung, P. ; McLoone, S.: MSC-clustering and forward stepwise regression for virtual metrology in highly correlated input spaces. In: *2012 23th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 2012, pp. 45–50

[127] Purwins, H. ; Barak, B. ; Nagi, A. ; Engel, R. ; Hoeckele, U. ; Kyek, A. ; Cherla, S. ; Lenz, B. ; Pfeifer, G. ; Weinzierl, K.: Regression Methods for Virtual Metrology of Layer Thickness in Chemical Vapor Deposition. In: *IEEE/ASME Transactions on Mechatronics* 19 (2014), Nr. 1, pp. 1–8

[128] Purwins, H. ; Nagi, A. ; Barak, B. ; Hoeckele, U. ; Kyek, A. ; Lenz, B. ; Pfeifer, G. ; Weinzierl, K.: Regression Methods for Prediction of PECVD Silicon Nitride Layer Thickness. In: *Proceedings of IEEE International Conference on Automation Science and Engineering*, 2011, pp. 387–392

[129] Pyle, D.: *Data Preparation for Data Mining.* Bd. 1. Morgan Kaufmann, 1999

[130] Qin, S. J. ; Cherry, G. ; Wang, J. ; Harrison, C. A.: Semiconductor manufacturing process control and monitoring: A fabwide framework. In: *Journal Process Control* 16 (2006), Nr. 3, pp. 179–191

[131] Ragnoli, E. ; McLoone, S. ; Lynn, S. ; Ringwood, J. ; McGearailt, N.: Identifying Key Process Characteristics and Predicting Etch Rate from High-Dimension Datasets. In:

Proceedings of the 2010 IEEE/SEMI Advanced Semiconductor Manufacturing Conference (ASMC), 2009, pp. 106–111

[132] Refaeilzadeh, P. ; Tang, L. ; Liu, H.: On comparison of feature selection algorithms. In: *Proceedings of AAAI Workshop on Evaluation Methods for Machine Learning II*, 2007, pp. 34–39

[133] Roberts, M. J. ; Russo, R.: *A Student's Guide to Analysis of Variance*. Routledge, 1999

[134] Roeder, G. ; Mattes, A. ; Pfeffer, M. ; Schellenberger, M. ; Pfitzner, L. ; Knapp, A. ; Mühlberger, H. ; Kyek, A. ; Lenz, B. ; Frisch, M. ; Bichlmeier, J. ; Leditzky, G. ; Ling, E. ; Zoia, S. ; Fazio, G.: Framework for Integration of Virtual Metrology and Predictive Maintenance. In: *Proceedings of the 23rd IEEE/SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 2012, pp. 288–293

[135] Roeder, G. ; Schellenberger, M. ; Pfitzner, L. ; Winzer, S. ; Jank, S.: Virtual Metrology for Prediction of Etch Depth in a Trench Etch Process. In: *Proceedings of the 24rd IEEE/SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 2013, pp. 326–331

[136] Roeder, G. ; Winzer, S. ; Schellenberger, M. ; Jank, S. ; Pfitzner, L.: Feasibility Evaluation of Virtual Metrology at the Example of a Trench Etch Process. In: *IEEE Transactions on Semiconductor Manufacturing* PP (2014), Nr. 99, pp. in print

[137] S., Liu ; Jia, C.-J. ; Ma, H.: A new Weighted Support Vector Machine with GA-based Parameter Selection. In: *Proceedings of 2005 International Conference on Machine Learning and Cybernetics* Bd. 7, 2005, pp. 4351–4355

[138] S., Pang ; Kasabov, N.: Inductive vs Transductive Inference, Global vs Local Models: SVM, TSVM and SVMT for Gene Expression Classification Problems. In: *IEEE International Joint Conference on Neural Networks*, 2004

[139] Saeys, Y. ; Inza, I. ; Larranaga, P.: A review of feature selection techniques in bioinformatics. In: *Bioinformatics* 23 (2007), Nr. 19, pp. 2507–2517

[140] Salomon, D.: *Data Compression: The Complete Reference*. Springer, 2004

[141] Sanchez-Marono, N. ; Alonso-Betanzos, A. ; Castillo, E.: A new Wrapper Method for Feature Subset Selection. In: *Proc. European Symp. on Artificial Neural Networks*, 2005, pp. 515–520

[142] Schellenberger, M. ; Roeder, G. ; Mattes, A. ; Pfeffer, M. ; Pfitzner, L. ; Knapp, A. ; Mühlberger, H. ; Bichlmeier, J. ; Valeanu, C. ; Kyek, A. ; Lenz, B. ; Frisch, M. ; Leditzky, G.: Developing a Framework for Virtual Metrology and Predictive Maintenance. In: *Future Fab International* 39 (2011), pp. 32–36

[143] Schirru, A. ; Pampuri, S. ; De Luca, C. ; De Nicolao, G.: Multilevel Kernel Methods for Virtual Metrology in Semiconductor Manufacturing. In: *IFAC World Congress*, 2011, pp. 11614–11621

[144] Schmitt, L. M.: Theory of genetic algorithms. In: *Theoretical Computer Science* 259 (2001), Nr. 1, pp. 1–61

[145] Sewell, M.: Feature Selection. In: *Available on http://machine-learning.Martinsewell.com* (2007)

[146] Sharma, D. ; Armer, H. ; Moyne, J.: A Comparison of Data Mining Methods for Yield Modeling Chamber Matching and Virtual Metrology Applications. In: *2012 23th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 2012

[147] Sheng, B.-Q. ; Pan, T.-H.: Virtual Metrology Algorithm for TFT-LCD Manufacutring Process. In: *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2011

[148] Shi, S. Y. M. ; Suganthan, P.N. ; Deb, K.: Multiclass Protein Fold Recognition using Multiobjective Evolutionary Algorithms. In: *Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2004

[149] Shlens, J.: A Tutorial on Principal Component Analysis. In: *Systems Neurobiology Laboratory, University of California at San Diego* (2005)

[150] Singh, M. P. ; Tiwari, R.: Correlation-based Attribute Selection using Genetic Algorithm. In: *International Journal of Computer Applications* 4 (2010), August, Nr. 8, pp. 28–34. – Published By Foundation of Computer Science

[151] Smola, A. J. ; Schölkopf, B.: A Tutorial on Support Vector Regression. In: *Statistics and Computing* 14 (2004), August, Nr. 3, pp. 199–222

[152] Spanos, J. C. ; Jula, P. ; Leachman, C. R.: The Economic Impact of Choosing offline, inline or insitu Metrology Deployment in Semiconductor Manufacturing. In: *Semiconductor Manufacturing Symposium, 2001 IEEE International*, 2001

[153] Stoppiglia, H. ; Dreyfus, G. ; Dubois, R. ; Oussar, Y.: Ranking a Random Feature for Variable and Feature Selection. In: *Journal of Machine Learning Research* 3 (2003), March, pp. 1399–1414

[154] Su, Y.-C. ; Cheng, F.-T. ; Huang, G.-H. ; Hung, M.-H. ; Yang, T.: A Quality Prognostics Scheme for Semiconductor and TFT-LCD Manufacturing Processes. In: *The 30th Annual Conference of the IEEE Industrial Electronics Society*, 2004, pp. 1972–1977

[155] Su, Y.-C. ; Lin, T.-H. ; Cheng, F.-T. ; Wu, W.-M.: Accuracy and Real-Time Considerations for Implementing Various Virtual Metrology Algorithms. In: *Transactions on Semiconductor Manufacturing* 21 (2008), August, Nr. 3

[156] Susto, G. A. ; Pampuri, S. ; Schirru, A. ; De Nicolao, G. ; McLoone, S. ; Beghi, A.: A Virtual Metrology System for Predicting CVD Thickness with Equipment Variables and Qualitative Clustering. In: *Proceedings of the 16th International Conference on Emerging Technologies and Factory Automation*, 2011

[157] Susto, G. A. ; Pampuri, S. ; Schirru, A. ; De Nicolao, G. ; McLoone, S. ; Beghi, A.: Automatic Control and Machine Learning for Semiconductor Manufacturing: Review and Challenges. In: *Proceedings of the 10th European Workshop on Advanced Control and Diagnosis (ACD 2012)*, 2012

[158] Susto, G. A. ; Schirru, A. ; Pampuri, S. ; De Nicolao, G. ; Beghi, A.: An Information-Theory and Virtual Metrology-based approach to Run-to-Run Semiconductor Manufacturing Control. In: *IEEE Conference on Automation Science and Engineering (CASE)*, 2012

[159] Suttorp, T. ; Igel, C.: Multi-Objective Optimization of Support Vector Machines. In: *Multi-Objective Machine Learning* Bd. 16. Springer Berlin Heidelberg, 2006, pp. 199–220

[160] Tay, F. E. H. ; Cao, L. J.: A comparative study of saliency analysis and genetic algorithm for feature selection in support vector machines. In: *Intelligent Data Analysis* 5 (2001), August, Nr. 3, pp. 191–209

[161] Therma-Wave, The Fab Productivity Enhancement C.: *Opti-Probe 3290, Film Thickness Measurement System, Specifications.* 3.2, 2002. – The documentation is confidential and dedicated to customer's internal use only, therefore not publicly accessible

[162] Thuselt, F.: *Physik der Halbleiterbauelemente.* Bd. 2. Springer, 2011

[163] Ting, K. M.: Discretization of continuous-valued Attributes and instance-based Learning / Basser Department of Computer Science, University of Sydney. 1994 (491). – Forschungsbericht

[164] Townsend-Weber, T. ; Kibler, D.: Instance-based Prediction of continuous values. In: *Working Notes of the AAAI-94 Workshop on Case-Based Reasoning*, 1994, pp. 30–35

[165] Tusher, Virginia G. ; Tibshirani, Robert ; Chu, Gilbert: Significance analysis of microarrays applied to the ionizing radiation response. In: *Proceedings of the National Academy of Sciences* 98 (2001), Nr. 9, pp. 5116–5121

[166] University, Stanford: Machine Learning / Stanford University. 2013. – Video Lectures

[167] Vapnik, V.: *The Nature of Statistical Learning Theory.* Springer, 1995

[168] Vapnik, V. N.: An Overview of Statistical Learning Theory. In: *IEEE Transactions on Neural Networks* 10 (1999), September, Nr. 5, pp. 988–999

[169] Venkata Naveen Kumar, N. ; Tsai, P.-F. ; Song, C. ; Wang, J. F. ; Mou, J.-I.: Iterative Backward Elimination PLSR: A novel PLS-based modeling technique to eliminate noise components for VM solutions. In: *e-Manufacturing & Design Collaboration Symposium (eMDC), 2013*, 2013, pp. 1–3

[170] Wang, L. ; Xu, G. ; Wang, J. ; Yang, S. ; Guo, L. ; Yan, W.: GA-SVM based feature selection and parameters optimization for BCI research. In: *Proceedings of the Seventh International Conference on Natural Computation*, 2011, pp. 580–583

[171] Wang, Y. ; Witten, I. H.: Induction of Model Trees for Predicting continuous Classes. In: *Proc European Conference on Machine Learning Poster Papers.* Prague, Czech Republic, 1997, pp. 128–137

[172] Weisstein, E. W.: *CRC Concise Encyclopedia of Mathematics.* 2nd edition. Chapman & Hall, 2003

[173] Weston, J. ; Mukherjee, S. ; Chapelle, O. ; Pontil, M. ; Poggio, T. ; Vapnik, V.: Feature Selection for SVMs. In: *Advances in Neural Information Processing Systems* 13 (2001), pp. 668–674

[174] Widmann, D. ; Mader, H. ; Friedrich, H.: *Technologie hochintegrierter Schaltungen.* Bd. 19. Springer, 1996

[175] Witten, I. H. ; Frank, E.: *Data Mining Practical Machine Learning Tools and Techniques.* 2nd Edition. Elsevier, 2005

[176] Wu, W.-M. ; Cheng, F.-T. ; Kong, F.-W.: A Dynamic-Moving-Window Scheme for Virtual-Metrology Model Refreshing. In: *Semiconductor Manufacturing, IEEE Transactions on* 25 (2012), Nr. 2, pp. 238–246

[177] Yang, J. ; Honavar, V.: Feature Subset Selection using a Genetic Algorithm. In: *IEEE Intelligent Systems* 13 (1998), April, Nr. 2, pp. 44–49

[178] Zeng, D. ; Spanos, C. J.: Virtual Metrology Modeling for Plasma Etch Operations. In: *IEEE Transactions on Semiconductor Manufacturing* 22 (2009), Nr. 4, pp. 419–431

[179] Zeng, D. ; Spanos, C. J. ; Tan, Y. ; Wang, T. ; Lin, C. ; Lo, H. ; Wang, J. ; Yu, C. H.: Virtual Metrology Modeling for Plasma Etch Operations. In: *2008 International Symposium on Semiconductor Manufacturing (ISSM)*, 2008, pp. 269–272