A Review of Data Mining Techniques for Failure Prediction in Continuous Casting

Authors

Rebecca Peters (left)
Ph.D. Student, School of Computing & Mathematics, University of South Wales, Treforest, Pontypridd, Wales
rebecca.peters@southwales.ac.uk

Penny Holborn (right)
Senior Lecturer, School of Computing & Mathematics, University of South Wales, Treforest, Pontypridd, Wales
penny.holborn@southwales.ac.uk

The Internet of Things (IoT) has seen an explosion in the development of sensing devices gathering and producing ubiquitous amounts of high-volume and high-velocity data contributing to the big data environment. Today, Industry 4.0 is driving advances in innovative technologies to help digitize manufacturing processes, enabling business to make more informed decisions. Sensors play a fundamental role in either monitoring or controlling the continuous casting process. Sensors continuously collect and transmit measurements and contain rich information about the condition of equipment. This paper explores the use of data mining techniques to generate failure prediction models where the pre-failure conditions are learnt from historical sensor data. The suitability of two failure detection approaches, anomaly detection and classification, are reviewed in collaboration with Tata Steel UK. Ultimately, this information will be applied in real time to predict future failures and to assist in optimizing the scheduling of maintenance.

With advances in technologies in recent years, the Internet of Things (IoT) has seen an explosion in the development of sensing devices gathering and producing sensory data to benefit a huge range of applications.[1] Whether it be in weather forecasting or to warn for early signs of medical problems, in today's society there is not an aspect of daily life that has not been impacted in some way by IoT. Sensing devices produce ubiquitous amounts of high-volume and high-velocity data contributing to the big data environment. The "3 Vs" are commonly used to explain the three key dimensions of big data: Volume, Velocity and Variety.[2] This was later extended to the "5 Vs" of big data, with the inclusion of two additional "Vs," Veracity and Value.[3] Hence, it is clear as to why there has been increased interest amongst the research community into investigating how best to extract value from multiple heterogeneous data sources in real time.[4]

Throughout the last 50 years, the steel manufacturing industry has become safer, cleaner, more cost-effective and more efficient as sensor technology has progressed. Today, Industry 4.0 is driving development in technologies such as Industrial Internet of Things (IIoT) to help digitize manufacturing processes.[5] The risk and consequences of performance degradation are significant, both in terms of productivity and safety. Hence, it is crucial to be able to diagnosis potential failures prior to the event occurrence. Prior to the big data era, existing research within the field of failure prediction relied on domain expertise and required complete data sets.[6] However, today's developments in artificial intelligence (AI) and machine-learning methodologies present techniques to handle incomplete and missing data.[7] For large steel manufacturing companies such as Tata Steel UK, sensors play a fundamental role in either monitoring or controlling the process or equipment. These sensing devices collect and transmit data on a continuous basis; however, there is limited current analysis of these signal measurements due to the challenges of utilizing big industrial data. These challenges are largely due to difficulties in being able to effectively manage and manipulate big data to be able to extract useful information.

There is extensive and well-documented research within the field of failure prediction. Existing research includes model-based approaches, signal processing and knowledge-based methods. Data mining includes traditional statistical methods and machine-learning techniques to extract insight from data.[8] However, there is limited research that considers extracting industrial big data and applying machine-learning techniques, specifically data mining, for predictive maintenance purpose. There is no known research that takes a holistic approach to applying the methods discussed within this paper to inform on machine maintenance for continuous casting of steel. Therefore, the aim of this paper is to review a range of machine-learning and data mining methodologies and to explore their potential application to failure detection in logged continuous caster sensor data.

Sensor-Based Failure Prediction

Sensor analytics is "the statistical analysis of data that is created by wired or wireless sensors."[9] Sensor analytics focuses on detecting patterns within sensor data. Some examples include alerting a medical practitioner of a patient's health, informing a data administrator about changes in an air conditioning system's functionality or the steel industry application considered within this paper — to highlight abnormalities in machine performance and predict potential failure of key continuous caster components. There are a range of techniques that can be applied to highlight fluctuations and deviation from what is considered "normal" or "healthy."

In traditional statistical analysis, time series data consists of a sequence of data points, which can be analyzed over some period: stock market values, daily temperatures, social media posts, etc.[9] Here, one of the key initial stages is to investigate the data to identify the change or shift points. These could be points that indicate a change in continuous casting segment performance, for example. Once these points have been identified, statistical parameters from the time series data are calculated. These parameters, for both the current time window and the previous window, are compared. Different techniques from the fields of traditional statistics, machine learning and data mining can be used to predict a shift. One approach could look to apply a Nearest Neighbor classifier[10] that compares each window in the training set to previously learned time series instances. The current window $Q[t - w,t]$, where Q is the continuous data stream, w is the window and t is the time series instance, is examined and if the time series distance falls below a certain threshold, then the engineers can be alerted.[11]

An individual data instance or record may consist of one or several features, e.g., the sensor reading, the confidence values and the sensor location. The sensor at a single time step is the record, and the sensor reading, the confidence value and the location are the features of the record. These features can be used to predict sensor readings. This terminology is widely recognized and originates from traditional statistical regression and analysis.[10] The data records also have labels; the labels "normal" and "anomalous" can be used to describe continuous caster machine activity or behavior.

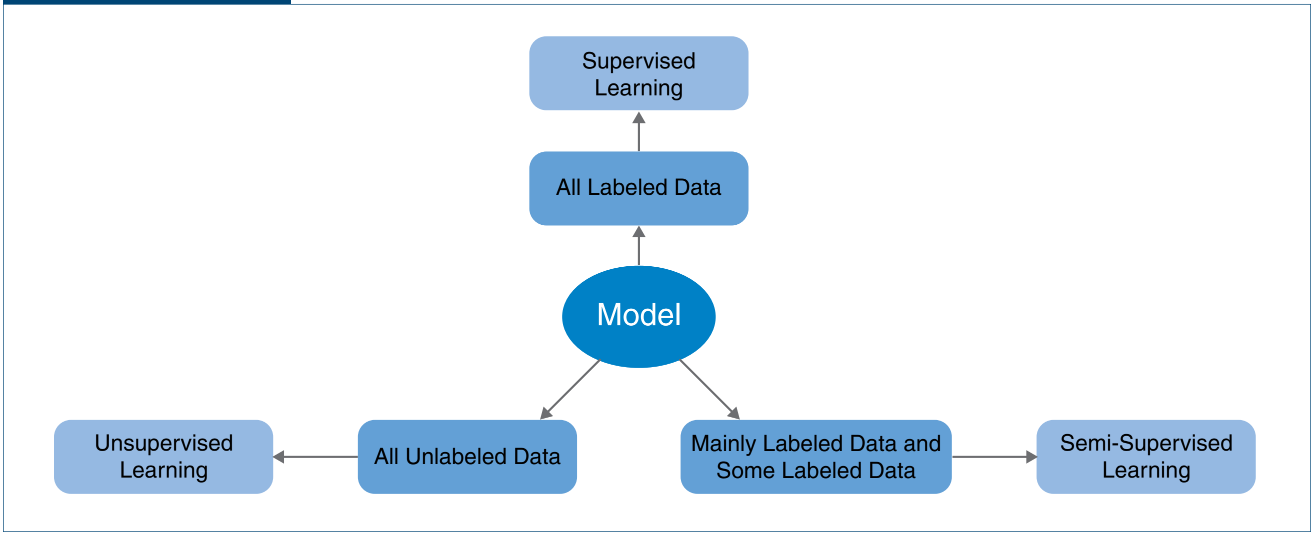
Sensor measurements hold rich information about a piece of equipment or a machine's pre-failure conditions. The pre-failure conditions and patterns within these sensor measurements can be "learnt" by sophisticated machine-learning algorithms to produce failure prediction models. There are two main proposed approaches to predicting failures from industrial big sensor data presented within this paper. These are failure prediction using classification techniques and failure prediction using anomaly detection methodologies. For anomaly detection methods, it is necessary to have labeled data to train a classifier to differentiate between anomalous and normal data instances. This approach of splitting the data into a training set and test set, for the algorithm to learn patterns in labeled instances for classification, is known as supervised learning. Unsupervised learning is usually considered for the case where the labels are unknown; these often include clustering algorithms. In addition to supervised and unsupervised learning approaches, there is a third learning type, known as semi-supervised learning. Semi-supervised approaches are generally considered for cases where only a small portion of the data set is labeled.[10] Fig. 1 illustrates these three learning approaches.

Assigning the data with the labels "anomalous" or "normal" is usually a manual task and would require a domain expert to confirm that the labels have been correctly allocated. Due to the size and complexity of industrial big data sets, semi-supervised or unsupervised learning approaches are recommended.

Sensor-Based Failure Prediction Literature

Methodologies for failure prediction were first presented in the 1960s. Within the literature, model-based approaches, signal processing and knowledge-based methods are three main contributors to feature detection for fault diagnosis. Knowledge-based methods utilize expert knowledge of key fault mechanisms and precise mathematical models. The approaches aim to predict imminent failure by comparing the actual residual signal to the predicted behavior. Parameter estimation[12] and expert systems[13] are examples of

Figure 1



Different types of learning approaches.

this type of approach. However, due to the harsh environment and the complex processes involved in steel manufacturing, it would be time-consuming to build and validate a model to achieve this. Methods for failure prediction utilizing signal processing focus on identifying faults in specific time and frequency windows. Signal operating features such as phase drift and amplitude can be successfully analyzed to determine a system's health.[14] For signal processing, data needs to be clean and complete. Industrial big data captured surrounding the continuous casting of steel is noisy and can contain inconsistencies resulting in incomplete data.

Advances in technology have led to the rise of "smart" objects (machines). These technological advances have driven a new paradigm shift in industrial production: Industry 4.0.[5] Fig. 2 outlines the four major shifts brought on by Industry 4.0 innovations, where IoT sensing devices have been a key contributor. Traditional steel manufacturing companies have been forced to invest in technology to increase productivity and to ensure that their market share is held in today's competitive market. With the exponential and continuous growth of big data, companies are keen to extract "value" from data to support data-driven decisions. There are a number of benefits in extracting value from industrial big data. These include predictive maintenance, improved machine performances, near-zero downtime and more.[15] When continuous casting vessels are off-line for maintenance, significant costs are accumulated, relating to a loss in steel slab production. Hence, a small reduction in maintenance frequency would have a large impact on both efficiency and machine performance. There is limited existing literature for sensor-based failure prediction using machine-learning techniques for predictive maintenance. Recent literature includes failure detection in semiconductor manufacturing,[16] tool wear prediction,[17] wind turbine gearbox failure,[18] etc. Both classification and anomaly detection have been identified as techniques that could be utilized to forecast abnormalities in logged sensor data captured during the continuous casting of steel. This paper focuses on evaluating these two key areas within the fields of data mining and machine learning. Different machine-learning algorithms require specific data pre-processing to first clean and convert raw data. Without pre-processing, the data failure prediction analysis is not feasible. Hence, the necessary stages of preparing big industrial data for machine learning failure prediction are discussed within this paper.

Pre-Processing Techniques for Big Industrial Data

Pre-processing industrial big data involves data cleaning (e.g., removing missing values, errors, outlier identification, etc.), data transformation (e.g., normalization, aggregation, feature engineering, etc.) and dimensionality reduction and integration (combining data from multiple data sources).[19] Data cleaning focuses on removing inconsistencies from the data to create a reliable and accurate data set for further analysis. For cases in which multiple data sources need to be integrated, the task of data cleaning increases significantly.[19] There are a number of tools available that can help support data cleaning, for example, OpenRefine, previously known as Google Refine.[20] Data cleaning and statistical analysis tools will be particularly useful to help detect and remove redundant sensor readings, remove misspellings from maintenance log data, and help refine slab quality and control data captured surrounding the continuous casting of steel. In summary, data pre-processing holds benefits in terms of data accuracy, completeness, consistency and uniformity.[21] Fig. 2 highlights these four key pre-processing stages of industrial big data for continuous casting of steel. These techniques are required to be performed prior to the deployment of machine-learning and data mining techniques.

Machine-Learning Approaches to Failure Prediction

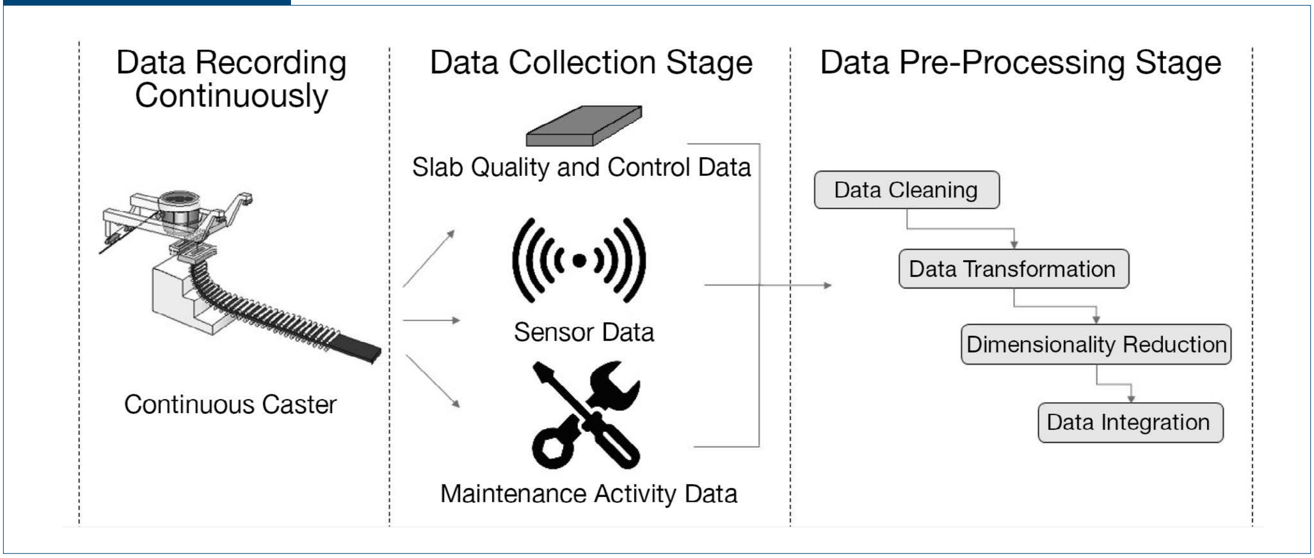
Machine learning has been successfully applied to a variety of tasks for a range of applications. These tasks include classification, prediction, understanding, and discovering data structure and anomaly detection.

Classification — Classification-based approaches require records to be classified as either normal or anomalous. Examples of techniques include Bayesian networks,[22] decision trees and rule-based classifiers,[23] random isolation forests,[18] and support vectors,[24] where rules can be defined by a domain expert or configured through machine learning. Most classification algorithms usually require the anomalies to be removed so that the algorithm can learn the patterns that reflect the normal operating conditions. The unseen data instance can then be classified into the category labeled normal behavior, otherwise the instance can be flagged and classified as anomalous behavior. Some classifiers can be trained on data that contains both instances labeled normal and anomalous. However, this approach requires the support of a domain expert to help decide on appropriate rules to classify future data successfully. Table 1 outlines different machine-learning classification techniques for failure prediction and highlights their advantages, disadvantages and examples of applications.

There are a number of advantages to machine-learning classification-based approaches for industrial failure prediction. To summarize the results of Table 1, the testing stage of the process is relatively quick in comparison to other approaches.[10] Unplanned downtime can have significant cost implications due to loss of production. Hence, for failure prediction in the continuous casting of steel, being able to deploy the model quickly and produce reliable results is vital. A further advantage of classification methods is that they can be applied in real time. Having data continuously fed into an automatically updating model would provide the required real-time diagnosis for continuous casting of steel. The model results could be used to populate a dashboard to inform casting engineers on the machine's current condition. There are also a range of powerful algorithms to choose from for classification. Having a large selection of techniques to apply helps to ensure that an optimal caster performance model can be achieved.

Despite the clear advantages to classification-based approaches for industrial failure prediction, there are also a number of limitations. The data needs to have been pre-assigned with labels indicating normal and anomalous behavior. This would require a casting domain expert's knowledge to validate the labeling of these instances, which would be a tedious and substantial task. Although classification algorithms have a quick testing phase, this is often counteracted by the significant time taken during the training phase.

Figure 2



Industrial big data pre-processing for continuous casting of steel.

Table 1

Machine-Learning Techniques for Industrial Failure Prediction

Classification technique	Description	Advantages	Disadvantages	Successful literature application
Bayesian networks	Used as a classifier to produce a posterior probability distribution of the class node given the value of other attributes. The class with the highest probability distribution is selected as the predictor.[25]	- Able to differentiate well between instances from different classes - Generally have a fast training phase in comparison to other algorithms	- For multi-classification, all the labels need to be pre-defined. Hence, anomalous class labels or multiple normal class labels are dependent on feasibility to label all normal classes. - Relies on human input to label the training records	Disk drive failures are usually infrequent events; however, when a disk driver fails, it demands large costs. Bayesian approaches can be applied to predict disk drive failures based on drive internal condition measurements.[22]
Decision trees and rule-based classifiers	Rule-based classifiers include techniques for classifying records using a series of "if [condition] then [conclusion]." Techniques aim to explain normal operating conditions for a system through a series of learning rules. A confidence interval can be used to determine the intervals that normalize the percentage of training data that needs to be evaluated correctly by the rule.[28]	- Use simple logic, hence are easy to interpret - Rule reasoning can be explained to highlight why an instance has been flagged as an anomaly - Able to quickly classify new records in real time	- Need to be carefully constructed to avoid infinite loops within rules - Contractions can arise when a new rule is added - For complex systems or processing, the rules can be complex and computationally costly - Require an extensive subset of normal training records to ensure that the rules explain normal behavior	In the rotary machine, industry roller bearing failure is extremely common. A range of statistical features such as minimum value, standard error, kurtosis, etc., can be extracted from vibrational signals. A fuzzy classifier can be built using a decision tree to automatically learn rules for fault diagnosis of roller bearings. The statistical features can help to indicate the difference between a healthy and faulty bearing status.[23]
Random forests/ isolation forests	An isolation forest is a random forest approach to outlier detection. By randomly selecting features and identifying values between maximum and minimum, observations are isolated. The length of the path is averaged over multiple decision trees (random forest); this provides a measure of normality for the decision function.[26]	- A well-studied machine-learning technique - Reduces data overfitting issues[10] - Normalization of features is not required	- Can be difficult to interpret in comparison to rule-based classifiers - As with other classification algorithms, depends on the identification of accurate and representative labels of different normality classes, which can be difficult	Both acoustic and vibration measurements provide information regarding the existence of gearbox faults. A random forest algorithm can be trained on a range of historical fault diagnostic experiments for the gearbox under different operating conditions. This algorithm presents a successful approach to monitoring gearbox condition.[18]
Support vector machines	A support vector machine is a discriminative classifier. The techniques are formally defined by a separating hyperplane. For failure prediction in a two-dimensional space, this algorithm aims to separate a plane into two parts where instances in each class normal and anomalous lie on either side.[10]	- One of the most efficient machine-learning techniques - Powerful for data separation - Scales well to high-dimensional data	- Lack of transparency of results - Interpretation relies heavily on graphical outputs to help visualize results - Requires a substantial training phase	Support vector machines can be applied to diagnose faults in low-speed bearings. Using acoustic emissions and accelerometer sensors under the same load at different speeds, support vector machines provide multi-class classification of faults.[24]

For the continuous casting of steel, the training stage would require pre-processing large amounts of data, which has potential to be computationally challenging. The primary aim of classification algorithms is to classify categories of normal behavior; anomalies are usually highlighted as a secondary goal. However, for the continuous casting of steel, the key focus is on identifying occurrences of anomalous behavior to predict potential failures — the aim being to aid predictive maintenance and warn casting engineers of earlier signs of machine degradation.

Anomaly Detection — Anomalies are points, items, events, outliers, or observations that fail to conform to some normal or expected behavior.[27] Anomaly detection focuses on identifying anomalous items; applications include fraud detection, medical diagnosis, errors in textual data, machine failure, etc. However, with advances in IoT, anomaly detection can be used

Table 2

Anomaly Detection Techniques for Industrial Failure Prediction

Classification technique	Description	Advantages	Disadvantages	Successful literature application
Markov chains	A Markov chain is constructed using historical data and the chain is used to determine the sequence of probabilities. For anomaly detection, the probability of a new sequence occurring can be determined and the rare sequences can be identified as anomalies.[32]	- Data-driven approach - Can achieve high accuracy and trade-offs with false positive and negatives	- Can be computationally expensive due to parameter estimation - Does not easily generalize - Not easy to scale to real-time operation	Markov chains can be applied to detect anomalous behavior in oil and gas pipeline pressure. The technique is applied to raw time series data, transferred into a Markov chain by a simple rule, and the transfer probability matrix of the Markov chain is calculated. The application considers two simple statistical features, mean and variance as the feature vector.[31]
K-nearest neighbors	K-nearest neighbors uses a distance measure to determine a suitable distance or similarity measure between instances. For anomaly detection, the approach relies on the assumption that normal records occur in dense neighborhoods, while anomalies occur far from their neighbors.[16]	- Data does not have to be pre-labeled as normal or anomalous - Data-driven assumptions regarding the distribution do not need to be considered.	- For big data, defining a distance measure can be challenging - Requires a large amount of computation for testing rather than during training.	Subsea valves are a crucial piece of equipment used to extract oil and natural gas. K-nearest neighbor algorithms can be applied to detect unusual valve behavior. This helps to assist maintenance engineers with condition-based monitoring.[33]
K-means clustering	K-means clustering originates from signal processing and methods of vector quantization. K-means clustering focuses on partitioning the data into k clusters where each cluster record belongs to the cluster with the nearest mean.[10]	- Can be applied in real time - Fast testing phase in comparison to other unsupervised learning methods	- For big data, defining a distance measure can be challenging - Approach is aimed at identifying cluster records that share similar characteristics not for identifying anomalies	K-means clustering has been successfully applied to fault detection in railway condition monitoring. K-means clustering was used to find appropriate parameters to detect and diagnose misalignments faults. The method was able to diagnose faults to a high degree of accuracy.[29]
Time series analysis	Time series analysis utilizes methods to analyze and forecast data over a particular time period. Time series methods for anomaly detection focus on forecasting a signal for some point and testing to see if this point deviates significantly from that forecasted. These points can then be identified as an anomaly.[32]	- If the assumptions regarding the distribution of the data are satisfied, then results can be accepted with confidence	- There is a risk that multivariate contextual anomalies can go undetected - Assumptions regarding the distribution need to be satisfied	A well-known time series analysis method of predicting future points in time series data is auto regressive integrated moving average. This method can be applied to predict vibration characteristics in rotating machinery. The approach was able to predict degradation trend of the machine and provide prognostics on the machine's health condition.[34]
Neural networks	Neural networks aim to simulate the operation of the human brain. They have been adopted into the field of anomaly detection to identify patterns in anomalous behavior.[32]	- Well suited to applications where there are multiple time series paired with each other	- Can be difficult to interpret - Trained on data to learn normal class/ classes	Neural networks can be applied to detect incipient faults in small and medium-sized induction motors. Measurements such as rotor speed and stator current are considered. Faults are detected to a high level of accuracy.[35]

to support use cases such as health monitoring and predictive maintenance. There are three main types of anomalies: point anomalies (lie outside a normal boundary region), collective anomalies (a subset or collection of anomalies) or contextual anomalies (points located closely are similar; this can be likened to anomalies found in time series data).[27] Contextual anomalies best describe the type of anomaly found in industrial sensor data. There are two main approaches for contextual anomaly detection:[28]

- To define the potential contexts and utilize point anomaly detection methods for each context.
- To model the normal structure in the data and reduce contextual anomalies to points anomalies. This includes identifying the anomalies within a training set and predicting the expected behavior based on previous records. An anomaly is highlighted if there is a considerable difference between the observed and expected value.

Machine-learning methods that consider anomaly detection for failure prediction in sensor data aim to learning normal clusters or groupings of data based on the spacing between sensor measurements. An anomaly score is calculated for each new measurement based on a distance measure assigned for pre-learned clusters.[29] Normal operating condition (NOC) models are generated based on the learnt historical data for detecting abnormal operating conditions during machine operation.[30]

Machine-learning methods for anomaly detection consider unsupervised and semi-supervised learning approaches for cases where there is no training data or the training data fails to meet the requirements for classification. A limitation of this approach is that there is no test set; therefore, it can be difficult to evaluate the performance. Table 2 outlines different anomaly detection methods for failure prediction and highlights their advantages, disadvantages and example applications.

There are a number of advantages to anomaly detection-based approaches for industrial failure prediction. To summarize the results of Table 2, the training stage of the process is relatively quick in comparison to supervised learning methods.[36] It is beneficial to have a fast training phase for streaming data captured surrounding the continuous casting process to ensure that the model can efficiently process the real-time data. If the training phase were to be computationally expensive, then it would be infeasible to retrain the model. Model retraining is required periodically to ensure accuracy is maintained. A further advantage of anomaly detection methods is that they can learn complex patterns; methods such as neural networks can use hidden layers between input and output to model intermediary representation of the data which other algorithms are not able to do. This is advantageous to failure prediction for the continuous casting of steel as the process controls and monitors multiple parameters that have complex patterns. The ability to model additional data representations could provide further insight into failures associated with the process. Unsupervised anomaly detection methods can highlight faults without first pre-assigning training instances as anomalous and normal. This is beneficial, as it would significantly reduce data preprocessing that would otherwise require a domain expert to validate a large set of training instances. However, a domain expert would still be required to validate the results of an unsupervised algorithm's output to ensure that the instances have been correctly identified.

However, similar to classification-based approaches for industrial failure prediction, there are also a number of limitations. For example, the K-nearest neighbors approach is memory-intensive and does not scale well to high-dimensional data. The continuous casting process considers high-dimensional data; hence this method would not be best suited to the application considered within this paper. A further disadvantage of unsupervised anomaly detection methods such as k-means cluster is that, for large data sets, identifying a meaningful distance function can be challenging. If the distance measure proves too challenging to identify, then alternate approaches may need to be explored for failure prediction in continuous casting. As mentioned previously, a further disadvantage to unsupervised methods is that although extensive input from the domain expert is not required, the process knowledge would still be required to validate the models output and also to fine tune critical levels.

Comparing Models and Deciding on the Best Approach for Failure Prediction in Continuous Casting of Steel

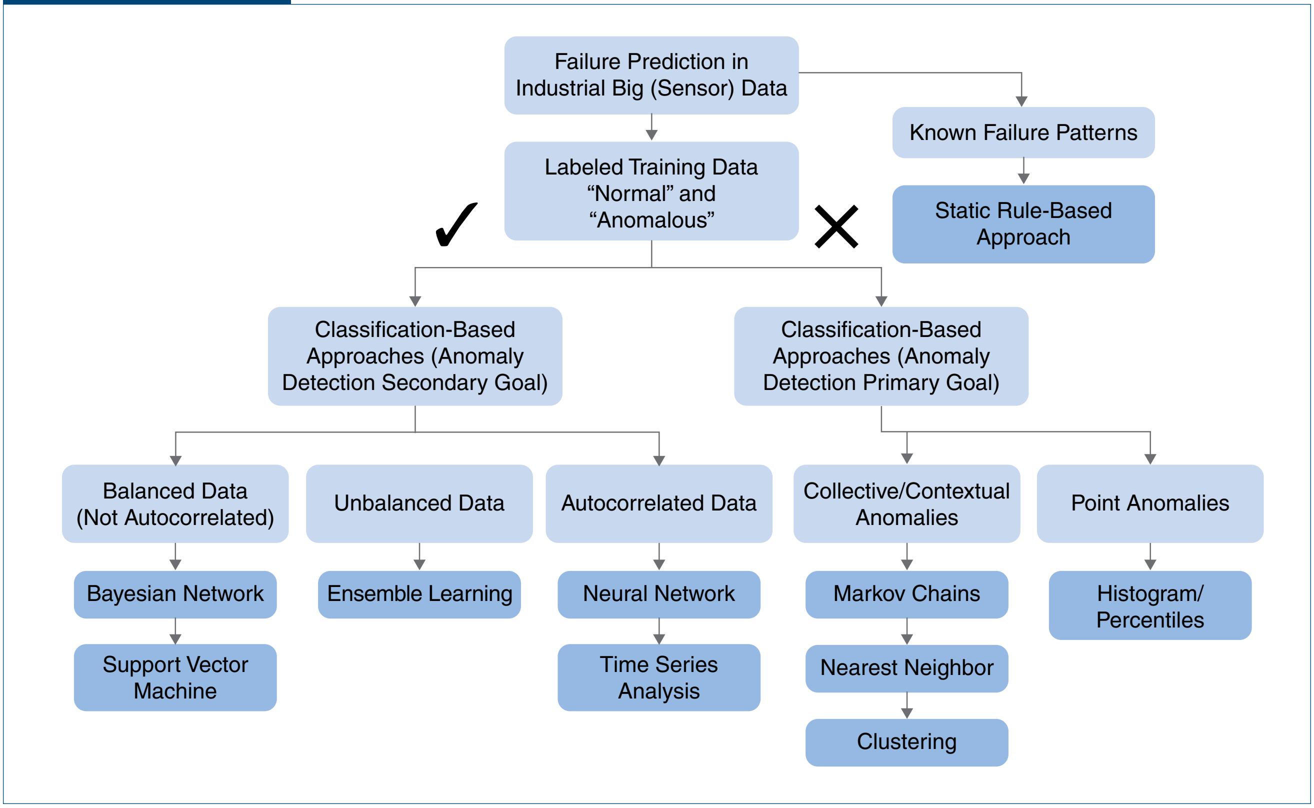
Fig. 3 illustrates a range of methodologies from fields of traditional statistical analysis, data mining and machine learning; in particular, suitable techniques to identify and forecast failures within industrial big sensor data. Fig. 3 shows that the decision to utilize either a classification-based approach or anomaly detection technique depends on the availability of labeled training data.

Classification-based approaches are applicable to predict failures if the training data is labeled, the number of anomalous and normal classes are balanced, and the data is not autocorrelated (i.e., one data point is not dependent on a previous data point). Where this is the case, rule-based random forests or support vector machines can be considered. However, for most IoT sources such as sensors, the data is often autocorrelated and unbalanced and anomalies are rare. Therefore, based on the size and complexity of data capture surrounding the continuous casting of steel, this study suggests that either an unsupervised or semi-supervised approach to failure prediction should be considered. This deduction of the suitability is largely due to the potential to only partially determine the health status of a continuous caster in historical data. These occurrences can be identified through aligning unplanned maintenance (known failures) events with fluctuations (deviations from what is considered normal behavior) in sensor measurements. Methods such as clustering provide the potential to build prognostic models for continuous casting of steel to forecast future vessel reliability.

Suggested System for Failure Prediction in Continuous Casting of Steel

Data mining technology brings new breakthroughs in failure prediction and active maintenance.[37]

Figure 3

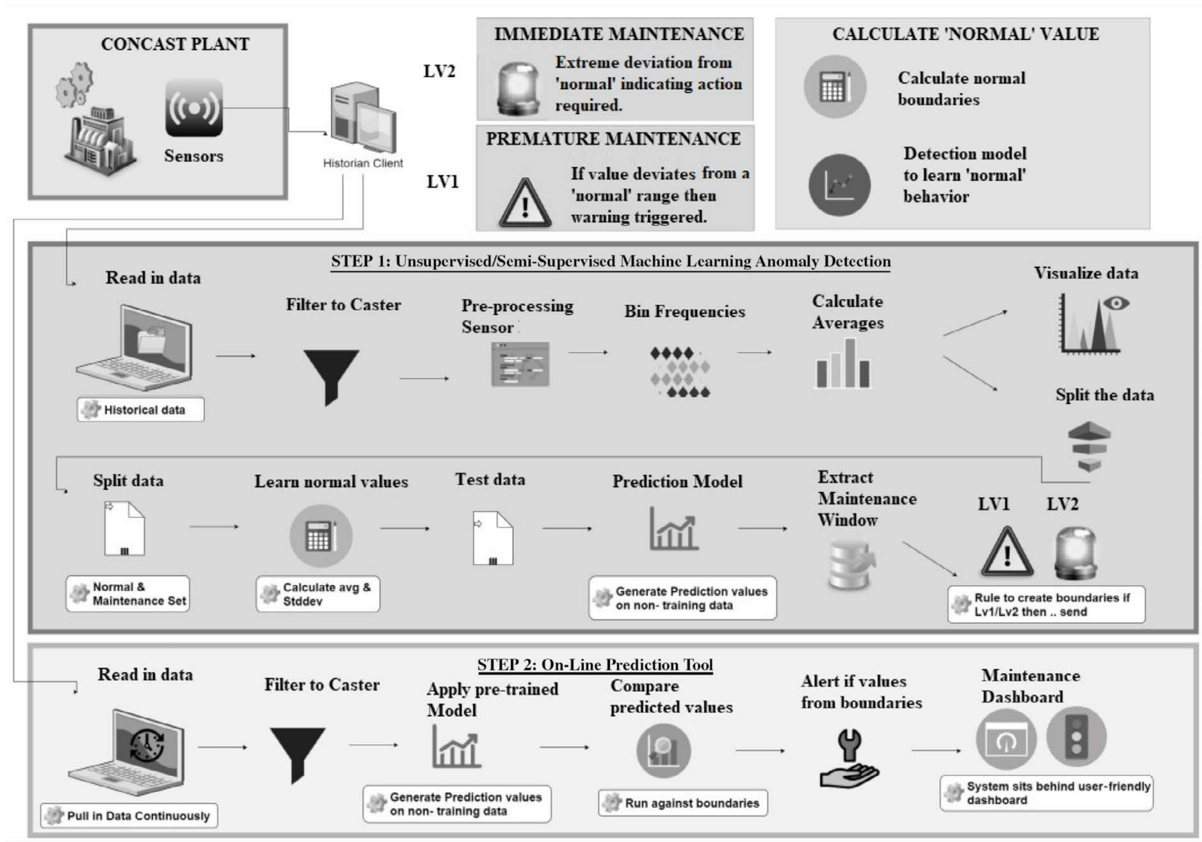


Failure prediction methods for industrial big sensor data.

Predicting failures before they occur could dramatically improve operational performance and support more informed decisions regarding the scheduling of maintenance. Fig. 4 presents the proposed predictive maintenance system for the continuous casting of steel slabs. For steel manufacturers, this system would help to reduce the occurrence of premature replacement of key casting components and minimize downtime.

Fig. 4 illustrates the various stages involved in developing and deploying an on-line tool to predict real-time failure in the continuous casting of steel. A range of data is recorded at the Concast plant, including bearing flow, pressure, temperature or speed. This information is held at the server and data analysis tools can be used to pull the data into a workflow. The first stage is to adequately prepare and pre-process the data by removing missing values, filtering the data to only include relevant information over a specific time period, merging multiple data sources and recordings, etc. Visualization and exploratory analyses are performed before the data is split into a training set and a test set. The training set contains mainly normal data that represent the continuous caster operating under healthy performance. The training set also includes some labeled occurrences of anomalous data for instances where the historical data can be aligned with previous emergency maintenance logs. The key maintenance windows are extracted; this knowledge can be utilized to warn engineers of the different levels of urgency of the potential maintenance required. By being able to monitor the assets in real time, a range of effective preventive measure can be automatically prescribed through work orders. The recommendations and details of what specific maintenance actions should be carried out will require expertise knowledge to define pre-established rules, tolerances or known critical levels. Once the model has been trained and tested, the model can be deployed in real time, where sensor measurements and additional data sources can be automatically compared to predicted values. Here, boundaries can also be altered and adjusted if necessary. If the actual data differs significantly from previously identified patterns of normal behavior, then the required alert level will be triggered. This insight can be delivered to the caster engineers through a traffic light maintenance dashboard. This type of dashboard has a number of benefits such as increased visibility, improved scheduling and effective asset management.

Figure 4



Failure prediction methods for industrial big sensor data.

Conclusion

This paper reviewed the suitability of two failure detection approaches, anomaly detection, and classification and their application to predict failures in the continuous casting of steels. It is clear that there is well-documented literature for the application of signal processing, knowledge-based and model-based approaches to failure prediction. However, there is less well-documented research to support the application of machine-learning techniques to extract value from industrial big data to predict failures and support predictive maintenance. Currently, there is no research to support a holistic approach to applying machine-learning methods to optimize maintenance scheduling of continuous casting of steel. Due to the continuous streaming of large volumes of data captured surrounding the continuous casting of steel process, this paper suggested an unsupervised or semi-supervised anomaly detection approach to identify failures. This is predominantly based on the computationally and time-expensive task of labeling training instances for classification.

To conclude, this paper provided insight to improve maintenance practices at large steel manufacturing companies such as Tata Steel UK. The ability to forecast and identy failures would be highly beneficial for a number of reasons. Failure detection methods present the potential to determine the frequency of failure, how often the failures occur and which components fail most frequently. The methods will provide further insight on the quality of available information and ease of detecting failures from the current data captured. It will also determine if there are specific failures for key components that can be more simply identified and forecasted, and if certain failure types are harder to predict. In addition, a further key benefit is the added knowledge on the impact the severity of the failure has on the continuous casting process. Future work will look to identify how different machine failures within continuous casting of steel influence both the productivity and the quality of steel produced.

References

1. J. Wan, D. Zhang, S. Zhao, L.T. Yang and J. Lloret, "Context-Aware Vehicular Cyber-Physical Systems With Cloud Support: Architecture, Challenges and Solutions," *IEEE Commun.,* Vol. 52, No. 8, August 2004, pp. 106–113.
2. Garnter, Big Data, 2001, https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
3. B. Mar, "Why Only One of the 5 Vs of Big Data Really Matters," IBM, 2015, http://www.ibmbigdatahub.com.
4. F. Chen, P. Deng, J. Wan, D. Zhang, A. Vasilakos and X. Rong, "Data Mining for the Internet of Things: Literature Review and Challenges," *Int. J. Distrib. Sensor Netw.,* Vol. 11, No. 8, April 2015, doi: 10.1155/2015/431047.
5. H. Kagermann, H. Albach, H. Meffert, A. Pinkwart and R. Reichwald, "Change Through Digitization — Value Creation in the Age of Industry 4.0," *Management of Permanent Change,* Springer Gabler, Wiesbaden, Germany, 2015.
6. X. Wu, X. Zhu, G. Wu and W. Ding, "Data Mining With Big Data," *IEEE Transactions on Knowledge and Data Engineering,* Vol. 26, No. 1, January 2014, pp. 97–107, doi: 10.1109/TKDE.2013.109.
7. Z. Ghahramani and M. Jordan, "Learning From Incomplete Data," Massachusetts Institute of Technology Artificial Intelligence laboratory and Centre for Biological and Computational Learning Department of Brain and Cognitive Sciences, A.I. Memo No. 1509, Paper No. 108, 1994.
8. SAS, "Machine Learning: What It Is and Why It Matters," SAS Institute Inc., https://www.sas.com/en_gb/insights/analytics/machine-learning.html.
9. M. Rouse, "Sensor Analytics," IoT Agenda, 2012, https://internetofthingsagenda.techtarget.com/definition/sensor-analytics.
10. T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* Springer, 2001.
11. G. Lesti and S. Spiegel, "A Sliding Window Filter for Time Series Streams," *Proceedings of the Workshop on IoT Large Scale Learning From Data Streams,* Skopje, Macedonia, Vol. 1958, 2017.
12. S. Cheng and M. Pecht, "Multivariate State Estimation Technique for Remaining Useful Life Prediction of Electronic Products," 2007.
13. H-J. Lee, B-S. Ahn and Y-M. Park, "A Fault Diagnosis Expert System for Distribution Substations," *IEEE Transactions on Power Delivery,* Vol. 15, No. 1, January 2000, pp. 92–97.
14. D. Ho and R.B. Randall, "Optimization of Bearing Diagnostic Techniques Using Simulated and Actual Bearing Fault Signals," *Mechanical Systems and Signal Processing,* Vol. 14, No. 5, 2000, pp. 763–788.
15. J. Lee, J. Ni, D. Djurdjanovic, H. Qiu and H. Liao, "Intelligent Prognostics Tools and e-Maintenance," *Computers in Industry,* Vol. 57, No. 6, 2006, pp. 476–489.
16. Q. P. He and J. Wang, "Fault Detection Using the K-Nearest Neighbor Rule for Semiconductor Manufacturing Processes," *IEEE Transactions on Semiconductor Manufacturing,* Vol. 20, No. 4, November 2007, pp. 345–354.
17. S.N. Huang, K.K. Tan, Y.S. Wong, C.W. de Silva, H.L. Goh and W.W. Tan, "Tool Wear Detection and Fault Diagnosis Based on Cutting Force Monitoring," *International Journal of Machine Tools and Manufacture,* Vol. 47, Nos. 3–4, 2007, pp. 444–451.
18. C. Li, R. Sanchez, G. Zurita, M. Cerrada, D. Cabrera and R. Vásquez, "Gearbox Fault Diagnosis Based on Deep Random Forest Fusion of Acoustic and Vibratory Signals," *Mechanical Systems and Signal Processing,* 2016, Vols. 76–77.
19. S. García, S. Ramírez-Gallego, J. Luengo, J. Manuel Benítez and F. Herrera, "Big Data Preprocessing: Methods and Prospects," 2016.
20. P. Larsson, "Evaluation of Open Source Data Cleaning Tools: Open Refine and Data Wrangler," University of Washington, 2013.
21. R. Blake and P. Mangiameli, "The Effects and Interactions of Data Quality and Problem Complexity on Data Mining," 2008, pp. 160–175.
22. G. Hamerly and C. Elkan, "Bayesian Approaches to Failure Prediction for Disk Drives," *Eighteenth International Conference on Machine Learning,* 2001, pp. 1–9.
23. V. Sugumaran, V. Muralidharan and K.I. Ramachandran, "Feature Selection Using Decision Tree and Classification Through Proximal Support Vector Machine for Fault Diagnostics of Roller Bearing," *Mechanical Systems and Signal Processing,* Vol. 21, No. 2, 2007, pp. 930–942.

A. Widodo, E.Y. Kim, J-D. Son, B-S. Yang, A.C.C. Tan, D-S. Gu, B-K. Choi and J. Mathew, "Fault Diagnosis of Low Speed Bearing Based on Relevance Vector Machine and Support Vector Machine," *Expert Systems with Applications,* Vol. 36, No. 3, Part 2, 2009.

24. D. Lunn, C. Jackson, N. Best, A. Thomas and D. Spiegelhalter, *The BUGS Book: A Practical Introduction to Bayesian Analysis,"* Chapman and Hall/CRC, New York, N.Y., USA, 2012.
25. F.T. Liu, K.M. Ting and Z. Zhou, "Isolation Forest," *2008 Eighth IEEE International Conference on Data Mining,* Pisa, 2008, pp. 413–422.
26. V. Chandola, A. Banerjee and V. Kumar, "Anomaly Detection: A Survey," *ACM Computing Surveys (CSUR),* Vol. 41, No. 3, 2009, doi: 10.1145/1541880.1541882.
27. P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández and E. Vázquez, "Anomaly-Based Network Intrusion Detection: Techniques, Systems and Challenges," *Computers & Security,* Vol. 28, No. 1–2, 2009, pp. 18–28.
28. T. Asada, C. Roberts and T. Koseki, "An Algorithm for Improved Performance of Railway Condition Monitoring Equipment: Alternating-Current Point Machine Case Study," *Transportation Research Part C: Emerging Technologies,* Vol. 30, 2013, pp. 81–92.
29. S. Manganaris, M. Christensen, D. Zerkle and K. Hermiz, "A Data Mining Analysis of RTID Alarms," *Computer Networks,* Vol. 34, No. 4, 2000, pp. 571–577.
30. S. Papavinasam, A. Doiron and R.W. Revie, "Model to Predict Internal Pitting Corrosion of Oil and Gas Pipelines," *Corrosion,* 2010.
31. D. Bhattacharyya and J. Kalita, *Network Anomaly Detection,* Chapman and Hall/CRC, New York, N.Y., USA, 2014.
32. M. Atif Qureshi, L. Miralles-Pechuán, J. Su, J. Payne and R. O'Malley, "VHI: Valve Health Identification for the Maintenance of Subsea Industrial Equipment," *Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2018, Lecture Notes in Computer Science,* Vol. 11053, 2019.
33. W. Wu, J. Hu and J. Zhang, "Prognostics of Machine Health Condition Using an Improved ARIMA-Based Prediction Method," *2007 2nd IEEE Conference on Industrial Electronics and Applications,* Harbin, 2007, pp. 1062–1067.
34. J.F. Martins, V. Ferno Pires and A.J. Pires, "Unsupervised Neural-Network-Based Algorithm for an On-Line Diagnosis of Three-Phase Induction Motor Stator Fault," *IEEE Transactions on Industrial Electronics,* Vol. 54, No. 1, February 2007, pp. 259–64, doi: 10.1109/TIE.2006.888790
35. T. Wuest, D. Weimer, C. Irgens and K-D. Thoben, "Machine Learning in Manufacturing: Advantages, Challenges and Applications," *Production & Manufacturing Research,* Vol. 4, No. 1, 2016, pp. 23–45, doi: 10.1080/21693277.2016.1192517.
36. J. Harding, M. Shahbaz, S. Srinivas and A. Kusiak, "Data Mining in Manufacturing: A Review," *ASME. J. Manuf. Sci. Eng.,* Vol. 128, No. 4, 2005, pp. 969–976, doi: 10.1115/1.2194554. ✦

This paper was presented at the 8th International SteelSim Conference, Toronto, Ont., Canada, and published in the Conference Proceedings.