Predictive System of Semiconductor Failures based on Machine Learning Approach

Yousef El Mourabit[1], Youssef El Habouz[2], Hicham Zougagh[3], Younes Wadiai[4]
TIAD Laboratory, Sciences and Technology Faculty, Sultan Moulay Slimane University, Beni Mellal, Morroco[1, 3, 4]
2Igdr Umr 6290 Cnrs- Rennes1 University, Rennes, France[2]

***Abstract*—Maintenance in manufacturing has been developed and researched in the last few decades at a very rapid rate. It's a major step in process control to build a decision tool that detects defects in equipment or processes as quickly as possible to maintain high process efficiencies. However, the high complexity of machines, and the increase in data available in almost all areas, makes research on improving the accuracy of fault detection via data-mining more and more challenging issue in this field. In our paper we present a new predictive model of semiconductor failures, based on machine learning approach, for predictive maintenance in industry 4.0. The framework of our model includes: Dataset and data acquisition, data preprocessing in three phases (over-sampling, data cleaning, and attribute reduction with principal component analysis (PCA) technique and CfsSubsetEval technique), data modeling, evaluation model and implementation model. We used SECOM dataset to develop four different models based on four algorithms (Naive Bayesian, C4.5 Decision tree, Multilayer perceptron (MLP), Support vector machine), according to the five metrics (True Positive rate, False Positive rate, Precision, F-Mesure and Accuracy). We implemented our new predictive model with 91, 95% of accuracy, as a new efficient predictive model of semiconductor failures.**

Keywords*—Machine learning; semiconductor; predictive maintenance; industry 4.0

I. Introduction

Currently, the industrial competition, the huge demand and the digital transformation have encouraged most industries to exploit and take advantage of the available technological tools. Many researches have proved the potential of the Artificial Intelligence (AI) for more efficiency and quality, reducing cost, and improving predictive maintenance services in manufacturing [1]. Nowadays, the industry is gradually developing towards what experts have called Industry 4.0, (The Fourth Industrial Revolution). This fact is strongly associated with the integration between digital and physical systems of production environments. This integration allows the collection of a huge amount of data by different equipment, located in many sectors of the factories [2]. Industry 4.0 is about performing tasks right on time, simultaneously, more efficiently, with more flexibility in a safer and respectful way. Likewise the industry 4.0 technologies integrate machines, products and people, allowing faster and more secure exchange of information [3]. The introduction of new technologies and new services associated with Industry 4.0 revolutionizes many industrial applications, and approaches, such as those in factories regarding automation and predictive industrial maintenance to create smarter work environments, in order to found opportunities for newcomers to de-liver innovative solutions that change business models.

Every year, a huge amount of data is collected by industrial systems, it contains precious information about processes and breakdown that occur in the production. In addition, analyzing and processing these data can show up valuable information and knowledge from system dynamics and manufacturing process [4]. Using various approaches based on data, it is possible to find illustrative results for strategic decision-making, providing advantages such as, increased production, machine fault reduction, maintenance cost reduction, among others [5] [6] [7]. The advantages above have strong relation with maintenance procedures. In manufacturing, equipment maintenance is a very important key, it affects the efficiency and operation time of equipment. As well, equipment faults need to be identified and solved, without production processes shutdown [8].

In literature, various groups and categories of maintenance management strategies can be found. Based on [9] [10], the maintenance procedures are classified as follows: Run-to-Failure (R2F) or Corrective maintenance, Preventive Maintenance (PvM) Time-based maintenance or Scheduled maintenance and Predictive Maintenance (PdM). PdM uses predictive tools to identify when maintenance actions are necessary. Therefore, it permit the early detection of failures by predictive tools using collected data with engineering approaches, statistical inference methods and machine learning techniques. To contribute to this challenge, in this paper we present a new predictive model of semiconductor failures, based on machine learning approach.

The aim of our work is to create a powerful predictive model of semiconductor failures that can predict future events to avoid failures. We used the SECOM dataset, after the preprocessing phases, we compare four predictive models based respectively on Naive Bayesian (NB), C4.5 decision tree, multilayer perceptron (MLP) , support vector machine (SVM) algorithms, in order to implement the most efficient and accurate model. According to several metrics (True Positive rate, false Positive rate, Precision, F-Measure and accuracy) we implemented a new efficient model based on MLP algorithm for predicting equipment failures during the wafer manufacturing process in the semiconductor industry, reached 91% of accuracy. The remainder of the paper is organized as follows: Section 2 introduces the related work. Section 3 presents our approach with detailed framework.

Section 4 details the experiment with a discussion of results. Section 6 concludes with future research directions.

II. Related Work

Artificial Intelligence transforms the traditional factory into a digital paradigm by increasing technological tools, real-time connectivity, and analytics capabilities. Moreover, data become a source of value to find illustrative results for strategic decision-making, in order to identify when maintenance actions are necessary, Hashemian and Bean [11] confirms that the few researches in the PdM area are due to the difficulty and complexity of implementing efficient PdM strategies in production environments. Also, the lack of use of Machine learning (ML) algorithms in PdM applications is related to availability of historical data in equipment failures, and especially, having professionals in the data science and ML field on the production line.

According to [12], Random forest (RF) is a supervised learning algorithm for regression tasks and classification. RF have shown more efficiency when the number of variables is larger than the number of samples. The main contributions of the Canizo work [13], are automation and scalability, also speed in data processing. Its results show an improvement of 5.54%, according to predictive accuracy, when compared to the Kusiak & Verma work [14]. The research developed by Su and Huang [15], presents a predictive fault detection system "HDPass", in order to perform hard disk drive faults. Using RF algorithm, the result presented by this work is promising, since it achieves 85% of accuracy. Authors used a type of SVM for regression purposes in [16]: Support Regression Vector (SVR). In this work, a modified regression kernel is presented to prognostic problems. In spite of that, the work does not perform any comparison between other ML methods. Results show that the proposed SVR model outperforms a standard SVR model.

Artificial intelligence, within, ML become a powerful tool for developing efficient predictive algorithms in various applications. ML approaches have the ability to deal with multivariate and high dimensional data in dynamic and complex environments [17]. Thus, ML offers powerful approaches for PdM applications. However, the efficiency of these applications depends on the adequate choice of the ML approach. Therefore, the aim of this paper is to present a new predictive model of semiconductor failures, based on the comparison of the most efficient machine learning algorithms (most used in PdM) according to various metrics.

III. Proposed Approach

Data preparation is the first critical phase in the development of a predictive model, it's an essential step aims converting various types and forms of data into an appropriate format, which is relevant to the predictive model based on machine learning. On the semiconductor manufacturing process, a huge amount of data is collected regularly during processing.

An experimental implementation was conducted to verify the efficiency and performance of the proposed failure prediction model by using SECOM dataset [18]. This dataset consists of 1567 data record and 591 attributes; it's collected from a semiconductor manufacturing process by monitoring the sensors and the process measurement point. Each record is a vector of 590 sensor measurements in addition to the data of the remaining feature were represented by Pass and Fail (label). Fig. 1 shows the proposed approach for generating a predictive model.

A. Oversampling Phase

The unbalanced distribution of data is a big challenge for standard learning algorithms. In SECOM dataset the number of successful tests is very important (1463 in-stances), compared to the number of failure tests, which is very infrequent (104 instances), this imbalance failure and success record, also the huge number of metrology data obtained from various sensors makes this dataset difficult to evaluate accurately. Therefore the forecasting model needs a data sampling method that can solve the imbalance of the records, for this we propose the sampling method [19].

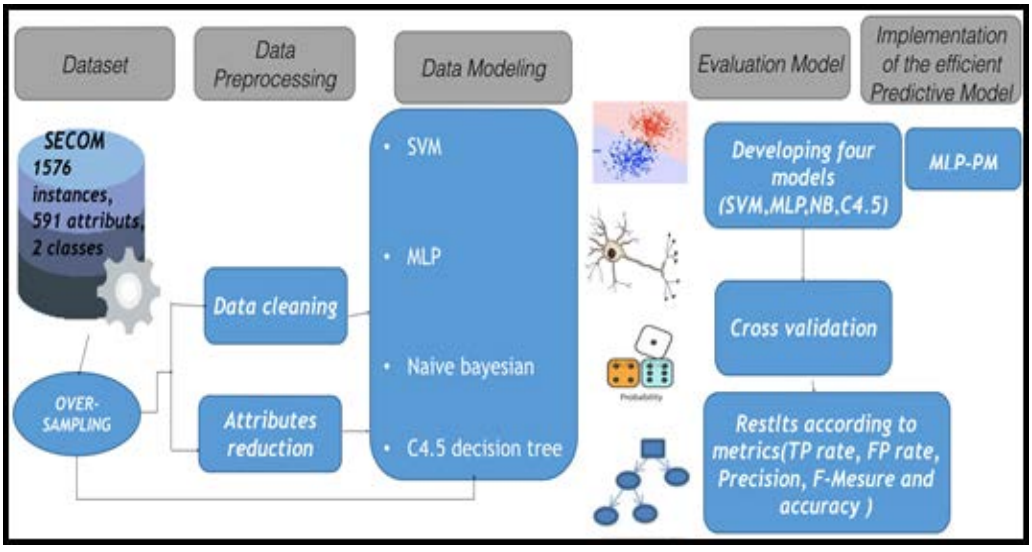

Fig. 1. Approach Architecture.

B. Data Cleaning Phase

Firstly we check the found value of each attribute, if the data seems to be unique value, i.e. the same value for all records, we remove this feature. Secondly we count in each column the missing data; if it reaches more than 55% we remove this attribute. We removed 158 attributes, and only kept 434 attributes.

C. Attributes Selection / Reduction Phase

A Huge amount of datasets are increasingly widespread in various disciplines. Characteristic selection or dimensionality reduction techniques are required to perform such datasets, also to improve prediction and computation performance, while preserving most information in the data. In this phase we used two methods to compare the results and implement the efficient one. Firstly we apply the CfsSubsetEval (Correlation-based Feature Selection) selection method [20], with the Best First search strategy that evaluates value of a subset of attributes according to the individual predictive capacity of each characteristic and the degree of redundancy.

The subsets of characteristics strongly correlated with the class while having a low intercorrelation are preferred. The result shows that just 17 attributes are considered, adding the label Pass / Fail, we obtain 19 attributes, we save the result as a separate dataset.

Secondly, we used principal component analysis (PCA) [21]. It aims to reduce the dimensionality of a dataset, and preserve as much as possible statistical information and variability. PCA geometrically projecting data onto lower dimensions named principal components (PCs), in order of finding the better summary of the data using a few number of PCs. Fig. 2 shows the correlation matrix of the used dataset.

After the standard scalar normalization, in order to normalize our set of features, we selected 168 best features to maintain approximately 95% of the accumulated variances as shown in Fig. 3.

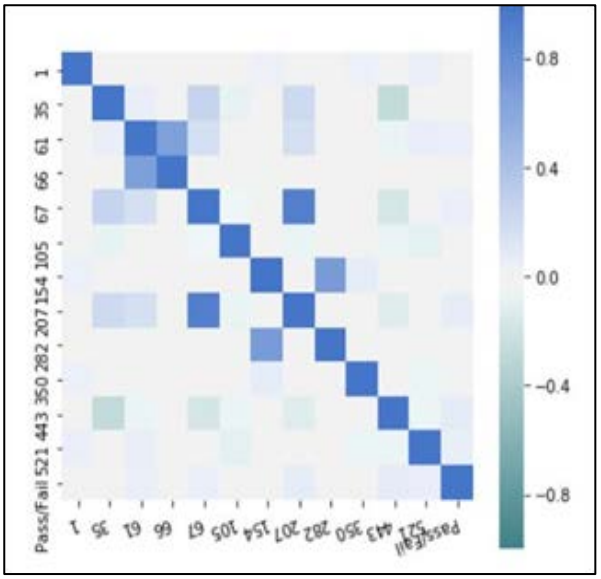

Fig. 2. Correlation Matrix of the used Dataset.

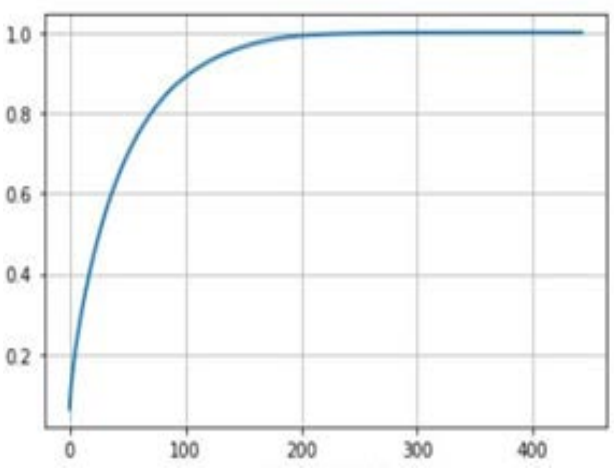

Fig. 3. Commutative Variance.

IV. Experiment and Results

After the preprocessing part, we performed a series of experiments in order to obtain the most efficient predictive model. Firstly, we used four different sets according to the data preprocessing phase: uncleaned dataset, cleaned dataset, cleaned dataset with attributes selected by the CfssubsetEval method, and cleaned attribute reduced by the PCA method.

Secondly, for every dataset, we applied four machine learning algorithms (SVM, NB, MLP, C4.5) [22], and perform the efficiency of these models based on five relevant metrics (TP Rate, FP rate, F-Measure, Precision, and accuracy) [23] [24]. Finally, we implement the most performant and efficient predictive model, based on MLP method using python environment. According to the results below, we can visualize the performance evaluation between the four different machine learning models, using four different dataset.

According to the results above (Fig. 4 to Fig. 8), it's clear that the using of a dataset with features reduction methods, improve significantly the accuracy of the four predictive models. Moreover, in this case, PCA method shows considerable performance compared to the CfssubsetEval method.

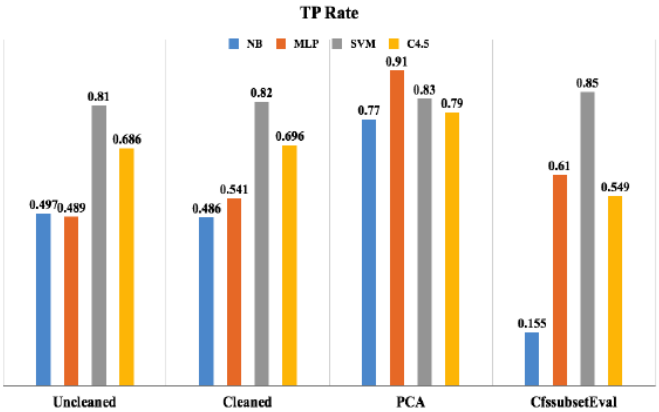

Fig. 4. TP Rate of Machine Learning Models according to the Four Datasets.

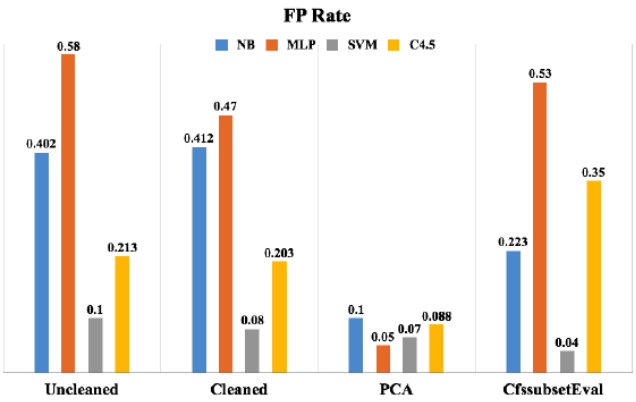

Fig. 5. FP Rate of Machine Learning Models according to the Four Datasets.

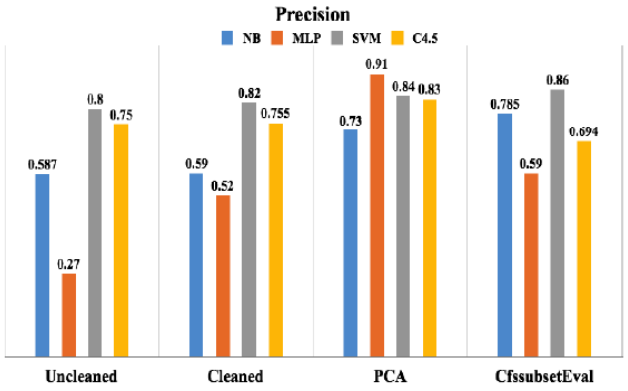

Fig. 6. Precision of Machine Learning Models according to the Four Datasets.

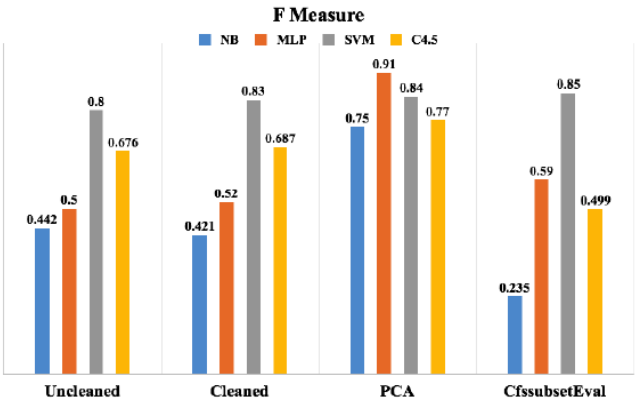

Fig. 7. F Mesure of Machine Learning Models according to the Four Datasets.

The highest rate of accuracy is obtained on MLP predictive model, using dataset with PCA features reduction method. It reached 91% of accuracy.

This presents the MLP model as the most efficient predictive model of semiconductor failures, which can predict future events to avoid failures.

In order to confirm this results, we implemented the MLP model on the python environment, then we obtained 91, 95% of accuracy.

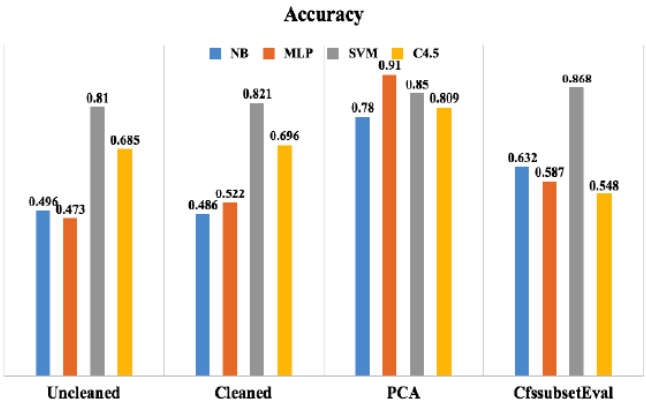

Fig. 8. Accuracy of Machine Learning Models according to the Four Datasets.

V. Conclusion

In order to create a new efficient predictive model of semiconductor failures based on machine learning, we designed and implemented four models based on the most used machine learning algorithms in this field, using SECOM dataset. Due to imbalance records of the success and failure examples in addition to the large amount of data we have proposed in the first part of preprocessing an oversampling method and during the cleaning phase, we have removed the attributes containing a unique value and the average 55% of missing values, and among these remaining features, we selected the most relevant features using CfssubsetEval methods and PCA method.

We performed a series of experiments from which we created four predictive models based on four machine learning algorithms (NB, SVM, MLP, C4.5). We implemented every model on four datasets (uncleaned dataset, cleaned dataset, and cleaned dataset with attributes selected by the CfssubsetEval method, cleaned and reduced with the PCA method.) Five metrics are used for efficiency evaluation (TP Rate, FP rate, F-Measure, Precision, and accuracy). Then, we developed the MLP predictive model on the python environment. The results shows that our predictive model is more efficient and performant, reached 91, 95% of accuracy. We report that data clearance and at-tribute reduction are critical steps in the data-mining process. Therefore we cannot ignore these phases, they require a considerable attention.

It is important to point that for dealing with maintenance events, PdM emerges as an efficient tool. With the Industry 4.0, PdM became gradually very promising. The employment of ML algorithms, for designing PdM applications leads to performant results with cost reduction of a PdM strategy in a factory.

In future works, we aim to use large and complex dataset with various labels, from different equipment on the factory, on real time, In order to identify other relevant features that impact the production line. Also, implement our model on real factory, and shows results on real time.

References

[1] J A. Shohin, X. Xun, L. Yuqian, et al. IoT-enabled smart appliances under industry 4.0: A case study. Advanced Engineering Informatics, 2020, vol. 43, p. 101043.

[2] T. Borgi, A. Hidri,B. Neef and M.S. Naceur. Data analytics for predictive maintenance of industrial robots. Interna-tional conference on advanced systems and electric technologies (IC_ASET) 2017 (pp. 412–417). IEEE.

[3] E. Rauch, C. Linder and P. Dallasega. Anthropocentric perspective of production before and within Indus-try 4.0. Computers & Industrial Engineering, 2020, vol. 139, p. 105644.

[4] T.P. Carvalho, F.A Soares, A. Fabrízzio, R. Vita et al. A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering, 2019, vol. 137, p. 106024.

[5] S. Biswal & G. R. Sabareesh, Design and development of a wind turbine test rig for condition monitoring studies. 2015 International conference on industrial instrumentation and control, ICIC 2015 (pp. 891–896). IEEE. (2015).

[6] R. S. Peres, R. A. Dionisio, P. Leitao and J. Barata. Idarts - Towards intelligent data analysis and real-time supervision for industry 4.0. Computers in Industry, 101, 138–146. (2018).

[7] E. Sezer, D. Romero, F. Guedea, M. MacChi and C. Emmanouilidis. An industry 4.0-enabled low cost predictive maintenance approach for SMEs: a use case applied to a cnc turning centre. IEEE International conference on engineering, technology and in- novation (ICE/ITMC) (pp. 1–8). IEEE. (2018).

[8] J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H. Abbas and A. V. Vasilakos. A manufacturing big data solution for active preventive maintenance. IEEE Transactions on Industrial Informatics, 13, 2039–2047. (2017).

[9] G. A. Susto, S. Member, A. Beghi and C. D. Luca. A predictive maintenance system for epitaxy processes based on filtering and prediction techniques. IEEE Transactions on Semiconductor Manufacturing, 25, 638–649. (2012).

[10] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone and A. Beghi. Machine learning for predictive maintenance: A multiple classifier approach. IEEE Transactions on Industrial Informatics, 11, 812–820. (2015).

[11] H. M. Hashemian and W. C. Bean. State-of-the-art predictive maintenance techniques*. IEEE Transactions on Instrumentation and Measurement, 60, 3480–3492. (2011).

[12] G. Biau, and E. Scornet. A random forest guided tour. TEST: An Official Journal of the Spanish Society of Statis-tics and Operations Research, 25, 197–227. (2016).

[13] M. Canizo, E. Onieva, A. Conde, S. Charramendieta and S. Trujillo. Real-time predictive maintenance for wind turbines using Big Data frameworks. IEEE International conference on prognostics and health management (ICPHM) (pp. 8). IEEE. (2017).

[14] A. Kusiak and A. Verma. Prediction of status patterns of wind turbines: A data- mining approach. Journal of Solar Energy Engineering, 133, 1–10. (2011).

[15] C. J. Su and S.F. Huang. Real-time big data analytics for hard disk drive predictive maintenance. Computers and Electrical Engineering, 71, 93–101. (2018).

[16] J. Mathew,M. Luo and C.K. Pang. Regression kernel for prognostics with support vector machines. IEEE Interna-tional conference on emerging technologies and factory automation (ETFA) (pp. 1–5). IEEE. (2017).

[17] T. Wuest, D. Weimer, C. Irgens and K.-D. Thoben. Machine learning in manu- facturing: Advantages, challenges, and applications. Production & Manufacturing Research, 4, 23–45, (2016).

[18] M. Salem, S. Taheri and J. Yuan. An experimental evaluation of fault diagnosis from imbalanced and incomplete data for smart semiconductor manufacturing. Big Data and Cognitive Computing, 2018, vol. 2, no 4, p. 30.

[19] J. Van Hulse and T. Khoshgoftaar.”Knowledge discovery from imbalanced and noisy data Data & Knowledge Engineering”, 2009.

[20] E.A. Bayrak, P. Kirci and T. Ensari. Performance Analysis of Machine Learning Algorithms and Feature Selection Methods on Hepatitis Disease. International Journal of Multidisciplinary Studies and Innovative Technolo-gies, 2019, vol. 3, no 2, p. 135-138.

[21] I.T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2016, vol. 374, no 2065, p. 20150202.

[22] I. Portugal, P. Alencar, and D. Cowan. The use of machine learning algorithms in recommender systems: A systematic review. Expert Systems with Applications, 2018, vol. 97, p. 205-227.

[23] D. Van Ravenzwaaij and J.P.A. Ioannidis. True and false positive rates for different criteria of evaluating statistical evidence from clinical trials. BMC medical research methodology, 2019, vol. 19, no 1, p. 218.

[24] S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023, 2018.