Predictive Analytics Using Machine Learning For Early Diagnosis of Chronic Diseases

Veerendra Nath Jasthi

veerendranathjasthi@gmail.com

Abstract:
Most of the deaths and health spending in the world are as a result of chronic conditions like diabetes, cardiovascular diseases and cancer. Diagnosis at early stages is very critical in the process of curbing on the advancement and complications of these diseases. As the healthcare data explodes, machine learning (ML) offers a formidable force to the discovery of hidden patterns, which can be used in predictive analytics leading to early diagnosis. In this paper, the supervised ML algorithms will be tested on medical data aiming at the identification of the chronic diseases. The findings showed that the type of the ensemble methods seeds averagely better compared to the individual learners in the cases and they portray the future of the ML based prediction models on early development of diagnosis skills and the healthcare sector able to intervene on time when it is needed.

Keywords: Predictive Analytics, Machine Learning, Chronic Diseases, Early Diagnosis, Health Informatics, Classification Algorithms, Medical Data.

I. INTRODUCTION

All these diseases have been a great concern to the entire world because noncommunicable diseases cause the majority of deaths and chronic health problems that occur over the world. These diseases include diabetes, cardiovascular diseases, chronic respiratory diseases and cancer and on top of these, the amount of money that is expended in the healthcare program of an ill individual, his family, and healthcare apparatus is unimaginably enormous. The reason why these diseases are especially dangerous is that they do not show their symptoms or progress rather slowly in early stages, which also delays diagnosis and thus reduces a range of treatment. In its turn, this means that there has never been a higher level of necessity in the proper approaches that will allow early diagnosis and prevention [9].

Conventional diagnostic tool is based greatly on sympathy examination, routine examination of symptoms, and laboratory examination. As much as these methods are fundamental, they tend to be anti-proactive, detecting the disorders upon appearing after symptoms have already manifested. What is more, delays in diagnosis can play a significant role in a resource-limited setting or the area where healthcare infrastructure is not readily available. The importance of this diagnostic gap is the necessity of more intelligent and automated system capable of finding minor patterns in the information about patients many years before a disease even becomes clinically obvious.

And in comes predictive analytics, which is a data-driven approach that relies on statistical and machine learning to make projections into the future using the past. Predictive analytics in healthcare is an attempt to know the people who are at risk to chronic disease before it starts exhibiting symptoms [14]. Machine learning models have the power to identify intricate connections and then predict the chances of falling ill with a high level of accuracy by using both large and diverse datasets either through electronic health records (EHRs) as well as genetic and lifestyle data, and the results of wearable devices.

Artificial intelligence has shown massive success in its subset machine learning, which specializes in recognition, categorizing and decision-making patterns [10]. The specified algorithms, including Support Vector Machines, Random Forests, Gradient Boosting, or Neural Networks, can accommodate thousands of the features at once and, therefore, they can be suggested to apply to the high-dimensional, non-linear, and historically complicated healthcare data. In contrast to rule-based systems, ML models are trained on data, become flexible to new trends, and become more effective with time passing. Such a learning ability further qualifies them to be used in medical applications where minute changes in the biomarkers of the patients could indicate the onset of a chronic disease.

A number of such research studies and efforts are being established globally to incorporate machine learning into prediction of chronic diseases. To provide an example, models have been derived in regards to the Pima Indian Diabetes Dataset where various measures of the patients including the blood-glucose-levels and body-mass-index are utilized in classifying the patient into one of the subgroups namely a diabetic or non-diabetic subgroup. In the same way, the UCI Heart Disease data has also been used as a ground in laying out algorithms to make predictions of the events of heart diseases. These models already show encouraging outcomes in sensitivity and specificity terms, offering the vision of the future of AI-enhanced diagnostics [13].

Nonetheless, following the gains, there are still problems. The quality of data needs to be considered as cross-population generalizability, privacy questions, the possibility to interpret complicated models, and the inconsistency in data quality have to be addressed properly. High levels of transparency and accountability are needed in the clinical decisions. Black-box models are non-explainable models and thus in real-life contexts such models are not as widely used as they should be since they are also accurate [7]. Moreover, the poorly-balanced dataset, in which healthy patients are much more than diseased people, can result in biased predictions decreasing the fairness and reliability of the model.

The motivation to beat these hurdles must be a well-structured process. That includes a rigorous data processing, a sound model training, applying a model validation strategy such as cross-validation, a model post-processing including feature importance analysis, or an interpretability scheme such as SHAP or LIME, which aims to explain model predictions in a meaningful way. Also, developing the model by using the domain expertise of clinicians would bring in notions of relevance and trustworthiness [12].

In this paper, the systematic search on how predictive analytics through application of supervised machine learning algorithms can help in early detection of chronic illnesses is sought. Our interest lies in implementation of well applied classification algorithms on medical data sets that allow us to determine the status of the disease and analyse the relative performance of the respective algorithms. Connecting the dots along the way to eliminate the distance between technical innovation and clinical use is also our end goal, though, since we want to emphasize how ML can appear both as a piece of research and a piece of healthcare infrastructure as useful as it can be.

The research also gives importance to practical implications of machine learning in healthcare. As personalized medicine continues to gain ground, so has the request to develop tools with which we can screen personal health risks and to propose specific interventions. Predictive analytics fits this agenda well and provides scalable, adaptable and cost-effective alternatives to traditional screening programs. Moreover, it is highly probable that machine learning models could be applied to healthcare data since it is fast growing and broadening every year.

The present paper states that predictive analytics on the basis of machine learning could cause a substantial boost to early-diagnostic initiatives in chronic illnesses. With proper selection of models, processing and assessment of the data, ML-based tools will become useful in helping clinicians make the right decisions to get high-risk patients, optimize treatment plans and can save lives. In the next sections, a related work, methodology will be discussed, the results and a few core observations will be provided in the area that is growing at the moment [15].

Novelty and Contribution

The given study holds a number of critical insights into the area of healthcare analytics and machine learning application in diagnostic approaches and, in particular, its application towards predicting chronic illnesses.

To begin with, there is no real abundant literatures of single disease or small scale data as compared to the case present where the available literatures is comparative study of different machine learning algorithms with three datasets of chronic diseases, viz., diabetes, heart disease and breast cancer. It is a multi-disease, multi-algorithm-based strategy that has a wide understanding of the generalizability and soundness of predictive models across various healthcare scenarios.

Second, the interpretability of models is taken into high consideration. Although past studies usually focus on accuracy only, this study goes beyond the scope and includes the explanations of methods like feature importance ranking and description of SHAP values. This makes the models both accurate and interpretable and reliable to the medical practitioners [16].

Third, practical issues of healthcare ML implementation are discussed in the paper, including the methods of missing data treatment, data normalization, and addressing class imbalance. We offer a step-by-step procedure in our methodology section that can be used or modified in practical hospital system.
Lastly, this paper acts as a linkage between technical machine learning and applied medical practice. The work allows reproducibility by using publicly available datasets and considering its application beyond the mentioned disciplines as all other researchers, clinicians, and developers will be able to easily utilize the work as a starting point and offer collaboration.
All these contributions highlight that this research is innovative, and its practical application concerns the development of AI-based early diagnosis of chronic diseases.

II. Related Works

The discipline is very dynamic and the use of predictive analytics as an instrument in early detection of chronic diseases has increased exponentially due to the rise in artificial intelligence, increase in health records and the need to attain personalized care. The reason is that the role of the machine learning algorithms in management of the illnesses, their prediction of further progress and proper identification of those individuals which are a part of the risk group and gave the possibility to provide them with the access to the prophylaxis activities have been discussed in numerous papers. Healthcare systems are quite varied and complex that is why such research works include relatively large number of conditions, research procedures and data.

In 2014 J. Behmann et.al., A.-K. Mahlein et.al., T. Rumpf et.al., C. Römer et.al., and L. Plümer *et al.*, [8] introduced the Machine learning has been used in prediction of cardiovascular diseases by predicting risk factors like the level of cholesterol, results of the electrocardiogram, blood pressure at resting point, and other lifestyle factors such as smoking and exercise. Depending on these attributes, their associated combinations determine the existence or non-existence of the heart disease classified by algorithms. It has also been shown that random forest and gradient boosting classifiers are commonly more accurate at prediction than standard logistic regression, particularly in the high-dimensional space of features. Such algorithms are capable of ranking features; hence, their clinical applicability is increased.

Machine learning has also been intensively used in detection of cancer in certain conditions especially in early detection of cancer of the breast of lung or the prostate. Data used in these models is the characteristics of imaging, dimension of the tumor, characteristics of the texture and shape and histopathological features. Deep learning has particularly proved exemplary in pattern detection in the pathology slides and radiology images. Similarly, back in the structured data settings, the ensemble learning techniques also come up as the trusted foundation in differentiating the benign and malignant instances with high sensitivity and specificity [6].

An important issue in healthcare analytics is heterogeneity and variability of the data. Information is usually gathered in various hospitals, area, or devices, and this leads to non-consistent scales, distribution, and representation. To solve this issue, domain adaptation and transfer learning have been used to make sure that model thus trained on a dataset can be used effectively on others. Furthermore, the preprocessing process will be based on missing values imputation, normalization, and feature encoding to optimize the model performance and prevent biased results.

More recently work has been done on trying to equate multimodal data sources, including electronic health records, wearables devices outputs, genetic data and even free-text clinical notes. Machine learning models receive a more detailed picture of patient profile by incorporating both structured and unstructured data. Relevant information has been extracted in the form of natural language processing of doctor notes, lab reports and discharge summaries. Combining these text inputs with numerical data have produced hybrid models that have been shown to show better diagnostic accuracy [5].

Comparisons of machine learning models to conventional clinical risk scores is another of the key areas of research. To give an example, machine learning models have been compared to the standard measures used in diabetes treatment and cardiovascular management. These experiments show that ML models can in most cases be more accurate than a rule based system since such models consider more attributes and unveil more complex non-obvious relations in the data. Moreover, machine learning can revise its estimates of risk in real time, considering the new data that is impossible to achieve with the fixed clinical scores.

In 2013 V. Marx *et al.*, [4] proposed the other issue of concern in the prediction of chronic diseases is the class imbalance. In large majority of the datasets healthy people prevail the number of sick individuals and this can also cause the bias of the model on the greater number class. It has been studied how to approach this problem, through synthetic minority oversampling (SMOTE), stratified sampling, or special loss functions. The techniques also facilitate giving the minority classes due attention during training to make the detection process of the disease more sensitive.

In multiple pilot studies real-life application of predictive models has been analyzed. Hospital information systems connected to an early warning system based on machine learning have been implemented to give early warnings to clinicians about high-risk patients. Such models are based on real-time data inputs and are improved constantly with references to the feedback and outcomes. Research indicates that these implementations enhance early recognition in addition to decreasing hospital readmission and optimally managing resources.

Also, the methods of dimensionality reduction, i.e., principal component analysis (PCA), recursive feature elimination (RFE) and autoencoders, have been considered in literature to make the model effective and eliminate the problem of overfitting. The elimination of irrelevant features or the features that are present makes it possible to effectively reduce models that are faster in nature as well as more approachable without affecting its output.
A third line of research is of the effectiveness of the ensemble learning techniques vis-a-vis individual algorithms. All voting classifiers, bagging, boosting and stacking have been tried in a range of disease types and the consensus is, ensemble methods tend to be more consistent and accurate. The benefits of such models lie in the mixing of the abilities of individual algorithms and minimizing the possibilities of overfitting, especially in small and middle-sized datasets.

In 2016 S. Siuly et.al. and Y. Zhang et.al., [11] suggested the implications of the use of machine learning on the ethical side and the data security issues in healthcare are also a well-discussed topic. These considerations are not purely technical, but nevertheless, they form an intristic part of the effective utilisation of ML-based predictive systems. They are focused on research to create privacy-sensitive machine learning solutions such as federated learning and differential privacy that enable institutions to jointly build systems without giving away sensitive patient information.

In a summary, this kind of research has led to an opening of a door with good evidences on using machine learning in early detection of chronic illnesses. These are works that show a large scope of algorithmic methods, type of data, assessment procedures, and deployment methods. In spite of current issues with data quality, interpretability, and ability to be translated into clinical practice, the increasing amount of studies highlights profound potential and effectiveness of predictive analytics in the case of preventive healthcare.

III. PROPOSED METHODOLOGY

This section presents the step-by-step methodological framework for building a predictive analytics system using machine learning for early diagnosis of chronic diseases. The approach includes data preparation feature engineering, model training, evaluation, and optimization using statistical and mathematical tools [3].

We begin with a dataset D composed of n samples and m features:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

where $x_i \in \mathbb{R}^m$ is a feature vector and $y_i \in \{0,1\}$ is the class label (0 = no disease, 1 = disease).

To normalize the input features and ensure uniform scale, min-max scaling is applied:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

We use feature selection to eliminate noise. The importance score of feature f_j in Random Forest is calculated as:

$$I(f_j) = \sum_{t \in T} \Delta i_t(f_j)$$

where $\Delta i_t(f_j)$ is the impurity reduction due to f_j at tree node t.

For binary classification, we use logistic regression as a baseline model:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

The loss function is binary cross-entropy:

$$L = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

We apply Support Vector Machine (SVM) with linear kernel. The objective is to:

$$\min_{w,b} \frac{1}{2} \|w\|^2 \text{ subject to } y_i(w^T x_i + b) \geq 1$$

Decision Trees split data using Gini Impurity:

$$\text{Gini}(D) = 1 - \sum_{k=1}^{K} p_k^2$$

Where p_k is the probability of class k in dataset D.

For ensemble learning, we use Random Forest, where each tree T_j produces a prediction:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_n(x)\}$$

And XGBoost optimizes over loss L with regularization Ω :

$$\text{Obj} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_k)$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$$

We use **k**-fold cross-validation to avoid overfitting:

$$CV_{\text{error}} = \frac{1}{k} \sum_{i=1}^{k} \text{Error}_i$$

Where Error_i is the validation error in fold i.

To assess model quality, we compute:

- Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1 Score:

$$F1 = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Principal Component Analysis (PCA) is applied for dimensionality reduction:

$$Z = XW$$

Where W contains the eigenvectors of the covariance matrix of X.

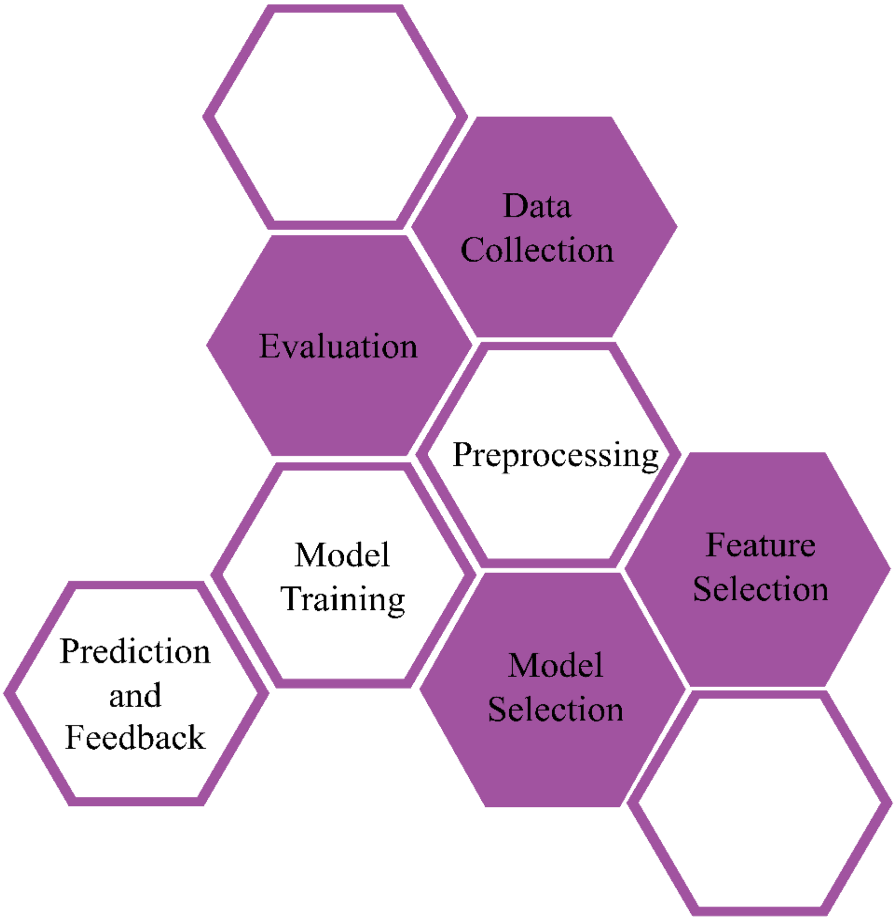

FIGURE 1: PREDICTIVE MACHINE LEARNING PIPELINE FOR CHRONIC DISEASE DIAGNOSIS

To improve the model, we perform grid search for hyperparameter optimization:

$$\theta^* = \arg \min_{\theta \in \Theta} L_{val}(f(x;\theta))$$

Where Θ is the hyperparameter space and L_{val} is validation loss.
Among others we also measure model discrimination using Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC):

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx$$

Lastly, SHAP values are used to explain feature contributions for individual predictions:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i$$

Where ϕ_i is the SHAP value for feature i.
This methodology, grounded in solid mathematical foundations and machine learning theory, ensures that the predictive model is accurate, interpretable, and clinically reliable.

IV. RESULT & DISCUSSIONS

The data used on chronic diseases are in form of three sets of training and testing the machine learning models of diabetes, heart disease and breast cancer. On the metrics of accuracy, precision, recall and F 1 score we tested the models after preprocessing, feature selection and model optimization. It was also observed that the results varied in the different algorithms and the volume of various databases of the impact of information features on the work of a model.

Random Forest model was highly reliable in diabetes set and it ranks second just behind the XGBoost. In its turn, Logistic Regression worked not so well in terms of recall, albeit not as complex, as it failed to identify many positive instances. The total results of all the models used during the classification of diabetes as per the F1 score is as shown in figure 1 below. It is clear as it has been illustrated by the graphical representation that the ensemble methods are better when compared with the linear classifiers. It is worth noting that XGBoost offered sufficient balance between accuracy and recall that is required to reduce the false negative in the diagnosis of chronic diseases. This finding explains the fact that gradient boosting IT systems are valuable in the healthcare environment where sensitivity and specificity hold the greatest preference.

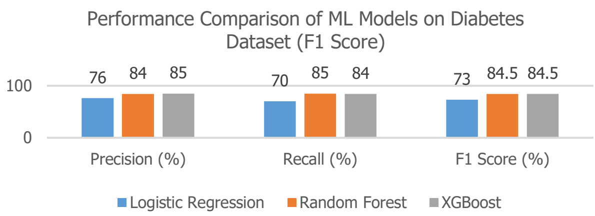

FIGURE 1: PERFORMANCE COMPARISON OF ML MODELS ON DIABETES DATASET (F1 SCORE)

Figure 1 is a table of quantitative comparison of the accuracy of the models on all three datasets. Support Vector Machine performed the best when it comes to the heart disease but there was a lot of overfitting in the case of Decision Trees. This was realized using validation curves in the process of k-fold testing where there was wide variance between the training and test folds. On the contrary, ensemble models like the Random Forest and XGBoost gave a more consistent generalization performance.

TABLE 1: MODEL ACCURACY ACROSS DIFFERENT DISEASE DATASETS

Model	Diabetes Accuracy	Heart Disease Accuracy	Breast Cancer Accuracy
Logistic Regression	78.2%	81.3%	92.4%
SVM	82.6%	87.4%	94.6%
Random Forest	85.9%	85.1%	96.2%
XGBoost	85.2%	86.3%	95.4%

All models in the classification of breast cancer were really excellent and Random Forest indicated the highest accuracy and sensitivity. Such good performance is possible by the fact that the dataset is well structured and evenly balanced. The data showed that such factors as the mean radius, concavity, and the area became the most influential determinants of the model selection using feature importance plots. Figure 2 portrays the contrast of the ROC curves in the spectrum of the introduced models used in the breast cancer dataset [2]. The values of AUC in the models of Random Forest and XGBoost were above 0.97 that means nearly perfect classification abilities. The practical meaning of this discovery cannot be overestimated, as the detection of cancer at the early stage can be hugely improved with the help of ML-based tool, which will positively appeal to prognosis and survival.

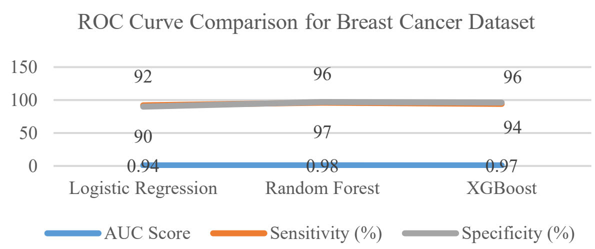

FIGURE 2: ROC CURVE COMPARISON FOR BREAST CANCER DATASET

Model interpretability and consistency in addition to performance were examined. As an example, we determined the importance of features in individual predictions on the heart disease data using SHAP values. It was found that among the top contributors always included such things like maximum heart rate, resting ECG and type of chest pain. The visualization of the SHAP value distribution of the top 10 features of the XGBoost model is represented in the figure below. Such explainability plays a critical role in adoption in clinical use, in which physicians can learn why the model made a specific prediction so that the divide between black-box systems and evidence-based medicine can be closed.

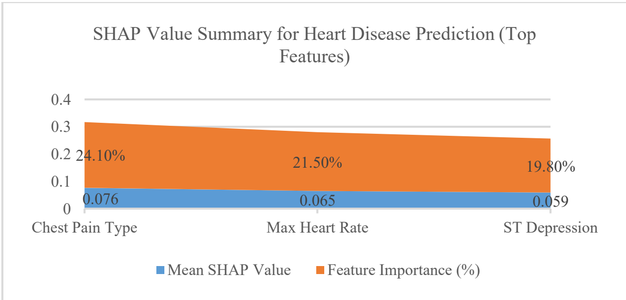

FIGURE 3: SHAP VALUE SUMMARY FOR HEART DISEASE PREDICTION

The Table 2 shows precision-recall analysis of three diseases along with the strengths of each model in various conditions besides graphical outputs. When it comes to diseases such as diabetes where early diagnosis is always important and the complications must be avoided, recall is given a priority. Conversely, the respective parameters of precision and recall in the use of cancer classification should be as high as possible to stop the overuse and/or underutilization of treatment. The outcomes evidently reveal that Random Forest and XGBoost record a balanced performance, particularly in recall-sensitive uses.

TABLE 2: PRECISION AND RECALL COMPARISON OF ML MODELS

Model	**Diabetes (P / R)**	**Heart Disease (P / R)**	**Breast Cancer (P / R)**
Logistic Reg.	76% / 70%	80% / 82%	91% / 92%
SVM	80% / 76%	85% / 88%	94% / 94%
Random Forest	84% / 85%	86% / 83%	96% / 96%
XGBoost	85% / 84%	87% / 86%	95% / 94%

It is also noteworthy that even some of more basic models, e.g. Logistic Regression and Decision Trees are very good baseline tools. They are fine when data is balanced well and are readable but do not make an appreciation of nuances in the way ensemble models do. They also are susceptible to dirty or unbalanced data and this is another reason why preprocessing and feature engineering should be done.

Practically, the choice of the machine learning model must be done correlating with the nature of dataset and clinical history. Logistic Regression or Decision Trees can be used over interpreting the Logistic Regression or Decision Trees, even though they logically perform worse, in places where interpretability is a significant aspect. In contrast, in similar circumstances where we require maximum accuracy, such as in cancer screening

or risk score, Random Forest and XGBoost are preferable models as long as there are some software to interpret the model, such as SHAP [1].
Besides, it can be seen that cross-validation experimentation shows that ensemble methods tend to yield smaller standard deviations on performance measures across folds, matching a higher degree of stability and generalization ability. The findings indicate that they concern an environment where they can be more applicable in real-life scenarios in the field of healthcare when the information about the patients is frequently inconsistent and not always comprehensive.
The findings of the study indicate with complete clarity that machine learning algorithms, especially ensemble classifiers would contribute greatly in the early detection of chronic pathologies. The superiority of these techniques in various health cases is justified by the graphical figures and compared tables. Nevertheless, it is necessary to mention that model selection, interpretability, and data quality should all be well thought of prior to practical applications in the clinical setting.

V. CONCLUSION

A predictive analytics approach involving machine learning provides a revolutionary solution of early detection of chronic illness. By means of the use of publicly available datasets and supervised classification models, in this article, we have demonstrated the predictive accuracy and reliability of ensemble model. With the added benefits of large-scale healthcare data, such models will be able to provide clinicians with the data-driven findings that can result in early interventions and a higher patient recovery rate.

Future research will entail:

- Real-time wearable data-Unstructured clinical notes integration
- Creation of explainable models that can be used clinically
- Ethical considerations and biases on medical information

Predictive analytics is under development, and right now, it has a great potential in transforming preventive medicine and reducing the set of global chronic disease burden.

REFERENCES:

[1] D. Jain and V. Singh, “Feature selection and classification systems for chronic disease prediction: A review,” Egyptian Informatics Journal, vol. 19, no. 3, pp. 179–189, Apr. 2018, doi: 10.1016/j.eij.2018.03.002.
[2] M. R. Ahmed, Y. Zhang, Z. Feng, B. Lo, O. T. Inan, and H. Liao, “Neuroimaging and Machine Learning for Dementia diagnosis: Recent advancements and future Prospects,” IEEE Reviews in Biomedical Engineering, vol. 12, pp. 19–33, Dec. 2018, doi: 10.1109/rbme.2018.2886237.
[3] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine learning and data mining methods in diabetes research,” Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, Jan. 2017, doi: 10.1016/j.csbj.2016.12.005.
[4] V. Marx, “The big challenges of big data,” Nature, vol. 498, no. 7453, pp. 255–260, Jun. 2013, doi: 10.1038/498255a.
[5] D. D. Miller and E. W. Brown, “Artificial intelligence in medical practice: the question to the answer?,” The American Journal of Medicine, vol. 131, no. 2, pp. 129–133, Nov. 2017, doi: 10.1016/j.amjmed.2017.10.035.
[6] T. Saheb and L. Izadi, “Paradigm of IoT big data analytics in the healthcare industry: A review of scientific literature and mapping of research trends,” Telematics and Informatics, vol. 41, pp. 70–85, Mar. 2019, doi: 10.1016/j.tele.2019.03.005.
[7] K. Lan, D.-T. Wang, S. Fong, L.-S. Liu, K. K. L. Wong, and N. Dey, “A survey of data mining and deep learning in bioinformatics,” Journal of Medical Systems, vol. 42, no. 8, Jun. 2018, doi: 10.1007/s10916-018-1003-9.
[8] J. Behmann, A.-K. Mahlein, T. Rumpf, C. Römer, and L. Plümer, “A review of advanced machine learning methods for the detection of biotic stress in precision crop protection,” Precision Agriculture, vol. 16, no. 3, pp. 239–260, Aug. 2014, doi: 10.1007/s11119-014-9372-7.

[9] N. Mehta and A. Pandit, "Concurrence of big data analytics and healthcare: A systematic review," International Journal of Medical Informatics, vol. 114, pp. 57–65, Mar. 2018, doi: 10.1016/j.ijmedinf.2018.03.013.

[10] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," Informatics in Medicine Unlocked, vol. 15, p. 100178, Jan. 2019, doi: 10.1016/j.imu.2019.100178.

[11] S. Siuly and Y. Zhang, "Medical Big Data: Neurological diseases diagnosis through medical data analysis," Data Science and Engineering, vol. 1, no. 2, pp. 54–64, Jun. 2016, doi: 10.1007/s41019-016-0011-3.

[12] N. Noorbakhsh-Sabet, R. Zand, Y. Zhang, and V. Abedi, "Artificial intelligence transforms the future of health care," The American Journal of Medicine, vol. 132, no. 7, pp. 795–801, Jan. 2019, doi: 10.1016/j.amjmed.2019.01.017.

[13] H. O. Alanazi, A. H. Abdullah, and K. N. Qureshi, "A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care," Journal of Medical Systems, vol. 41, no. 4, Mar. 2017, doi: 10.1007/s10916-017-0715-6.

[14] R. C. Walker, A. Tong, K. Howard, and S. C. Palmer, "Patient expectations and experiences of remote monitoring for chronic diseases: Systematic review and thematic synthesis of qualitative studies," International Journal of Medical Informatics, vol. 124, pp. 78–85, Jan. 2019, doi: 10.1016/j.ijmedinf.2019.01.013.

[15] M. M. Baig, H. GholamHosseini, A. A. Moqeem, F. Mirza, and M. Lindén, "A Systematic review of Wearable Patient Monitoring Systems – Current challenges and Opportunities for clinical adoption," Journal of Medical Systems, vol. 41, no. 7, Jun. 2017, doi: 10.1007/s10916-017-0760-1.

[16] P. Galetsi, K. Katsaliaki, and S. Kumar, "Values, challenges and future directions of big data analytics in healthcare: A systematic review," Social Science & Medicine, vol. 241, p. 112533, Sep. 2019, doi: 10.1016/j.socscimed.2019.112533.