Predictive Analytics for Multiple Diseases Using Machine Learning

Vijayalakshmi M., Shiva Subramanian and Harsh Jain
Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, India

Keywords: Chronic Diseases, Medicare, Health Care Management Data Analytics, Machine Learning, Supervised Learning, Disease Prediction, Support Vector Machine (SVM), Logistic Regression, Early Detection, Patient Outcomes, Predictive Analytics.

Abstract: The prevalence of chronic diseases is increasing among Medicare patients, and there is a need for innovative ways of healthcare management. Medical practitioners are commonly swamped by the large quantities of information involved in the analysis so that it becomes an uphill task to interpret symptoms and diseases within time. Utilization of supervised ML algorithm has proved its effectiveness for diagnosis purposes to diseases and enables medical professionals to detect risky conditions in an early stage. The goal of the project is to predict the likelihood of a variety of diseases in the Medicare population using data analytics and machine learning. We will identify patterns and risk factors associated with multiple diseases through preprocessing and enhancement of available data. More advanced ML algorithms will be leveraged to create prediction models: SVM for the prediction of both diabetes and Parkinson's disease and logistic regression for heart disease. Feeding labeled input data into the algorithms during the training process will help learn correlations between feature and disease correlations. When predicting, the models will then be tested on an independent data set to establish how accurately they are able to predict outputs and call out potential issues for fine-tuning if needed. Such insights help healthcare providers identify a patient issue earlier and intervene appropriately. This, of course, improves patient outcomes while controlling health costs. This project depicts the promise of predictive analytics with respect to improving patient care under Medicare through creating personalized and proactive healthcare solutions.

1 INTRODUCTION

The prevalence of chronic disease in Medicare patients has been on the rise and it has become a challenge for healthcare management, thus calling for new ways of facilitating improved detection and management of such diseases. Health care professionals deal with large amounts of information related to patients. In that context, proper diagnosis of symptoms and detection of diseases at an early stage has become difficult. Traditional procedures fail to meet the volume and complexity of this information, so advanced techniques have to be employed to make health care delivery effective. In such a context, supervised machine learning algorithms are becoming promising solutions in improving disease diagnosis and early detection. The algorithms rely heavily on datasets in recognizing patterns and at-risk factors of various diseases, thereby supporting medical practitioners to make informed choices. This project is designed in a way to utilize the strength of data analytics and machine learning for predicting the onset of chronic diseases in the Medicare population. The existing state-of-the-art ML models for disease prediction have the issues of poor data quality, limited generalizability, overfitting, and strong feature selection that may lead to nonreliable predictions, so this project is focused on promoting better preprocessing of data, application of multiple ML algorithms, and using validation with the use of various datasets to improve accuracy and generalizability together with better interpretability. SVM algorithms will be used to predict diabetes and Parkinson's disease, while logistic regression will be used in predicting heart disease. Once these models are put through training with labeled patient data, the algorithms will have learned to recognize patterns between features and their corresponding disease. These models will then be tested using another independent set of data and, therefore, their accuracy will be checked for further verification about potential problems. This manner, the results that are going to come out would be

M., V., Subramanian, S. and Jain, H.
Predictive Analytics for Multiple Diseases Using Machine Learning.
DOI: 10.5220/0013923300004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 5*, pages 92-99
ISBN: 978-989-758-777-1

reliable enough to make necessary adjustments. The results expected from this project are going to guide the providers of healthcare into earlier identification and intervention for chronic diseases with the aim of rising patient outcomes and lowering health care costs in facilities. This paper focuses on the integration of predictive analytics in Medicare patient care. Thus, it helps in giving personalized and proactive healthcare solutions, transforming the management of chronic diseases.

State-of-the-art machine learning-based disease prediction systems face substantial challenges like dependency on the quality of the dataset, challenges in feature selection, and origination of bias through the use of a single algorithm. These issues might lead to false predictions and the defeat of precocity of diseases. To counter these issues, this project aims at designing a multi-algorithm machine learning system that utilizes a variety of supervised learning models in order to have an increased level of accuracy pertaining to predictiveness and to take away the bias. Data preprocessing along with advanced techniques of feature selection is also crucial in order to improve quality to a large extent. These need to be expanded to include different demographics to make the models more generalizable and less likely to overfit. Evaluation of the learning models after training will also increase the reliability of predictions. Lastly, it will evaluate and compare the performance of different machine learning models applied in handling various diseases, including heart disease, diabetes, and Parkinson's disease. By addressing these objectives, the project aims to offer a strong framework for more accurate and reliable disease prediction.

2 LITERATURE SURVEY

The inclusion of machine learning (ML) in healthcare for disease prediction has highlighted significant challenges and limitations in current systems. A major issue is the dependency on data quality. Poor data, characterized by missing entries, inconsistencies, and class imbalances, often results in unreliable predictions. Kourentzes et al. and Lipton et al. demonstrated how inadequate data quality adversely impacts model accuracy, especially in predicting cardiovascular disease mortality risks. Beam and Kohane also emphasize that data preprocessing is critical for healthcare applications to address these issues effectively.

Another key challenge is effective feature selection. Guyon and Elisseeff identified that inefficient methods tend to overfit, capturing noise rather than meaningful patterns. Prominent techniques like Recursive Feature Elimination and LASSO (Zou and Hastie) have proven useful but often lack consistency across varied datasets. Miotto et al. highlighted the need for dimensionality reduction techniques to avoid overfitting while preserving model effectiveness.

Algorithmic bias is another significant concern. Obermeyer et al. observed that models trained on homogeneous datasets struggle to generalize, thereby exacerbating health disparities among demographic groups. To address this, multi-algorithm systems leveraging strengths from multiple supervised learning models have emerged as a promising solution. For instance, Hodge and Austin demonstrated how ensemble methods could enhance accuracy and reduce bias, particularly in predicting diabetes and heart disease (Chakraborty et al.,). Esteva et al. explored how deep learning models could reduce bias by integrating diverse datasets.

Advanced preprocessing and feature engineering techniques also play a critical role in improving model performance. Sun et al. conducted systematic reviews highlighting the importance of preprocessing in enhancing generalization and minimizing overfitting. Topol emphasized the importance of high-quality preprocessing for AI-driven healthcare solutions. Furthermore, Wang and Preininger stressed the need for diverse datasets to ensure robust and unbiased model training.

Model evaluation is equally crucial. Good post-training evaluation techniques, such as cross-validation, ensure performance optimization and reliability (Varma & Simon). Liu et al. demonstrated that systematic evaluations could match or outperform human-level diagnostic accuracy in certain medical domains.

Recent advances in machine learning have shown immense potential for addressing these challenges. Liang et al. applied AI to pediatric disease diagnostics, achieving high accuracy and interpretability. Shickel et al. highlighted the growing role of deep learning in analyzing electronic health records (EHRs), enabling better predictions for chronic diseases. Rajkomar et al. explored how ML could revolutionize patient care by integrating precision medicine and scalable analytics.

Additionally, Esteva et al. and Litjens et al. demonstrated how deep learning models, such as convolutional neural networks, could achieve dermatologist-level performance in detecting skin conditions and medical imaging, respectively. Advances in these areas pave the way for

implementing predictive analytics across diverse healthcare domains.

Taken together, these insights underscore the need for creative solutions to improve ML-based disease prediction systems. They also highlight the importance of data diversity, advanced preprocessing, multi-algorithm approaches, and rigorous model evaluation to ensure reliable and generalizable predictions.

3 WORKING PRINCIPLE

The multi-algorithm approach shall use several supervised machine learning models in order to enhance the accuracy of disease prediction with balancing the type of bias existing with single-algorithm-based approaches. Advanced preprocessing techniques come into effect initially for cleansing and preparing the dataset while maintaining the quality by determining such issues as missing values and inconsistencies. Then feature selection processes are integrated to identify and rank features important to the predictive performance of the models.

During training, cleaned and preprocessed data are fed into sets of ML algorithms, which include SVM for diabetes and Parkinson's disease prediction models and logistic regression for the heart disease prediction model. The models are trained with labeled data such that it learns the pattern and correlation between features of the patients and the outcome of their diseases. Lastly, the trained models were validated on separate sets of data that did not take part during the training process. This validation would help in establishing the correctness of the models besides generalizing over other patient populations.
After validation, the models are integrated in user-friendly interfaces based on Streamlit that enable users to input a disease they want to check. The application captures the input from the user performs the needed validations and then processes it through the appropriately trained model. As a result, the returned prediction results along with their respective confidence levels are returned to the user for enabling decision making pertinent to the health area. This is a structured approach that will target the provision of a strong and reliable predictive analytics tool for disease detection in early stage development to improve patient outcome results while enhancing healthcare delivery efficiencies

4 METHODOLOGY

4.1 System Design

In the system design, the choice of Streamlet framework as a means of building the user interface is what makes up the major elements of the software, coupled with pre-trained models of machine learning like Support Vector Machines and logistic regression, as well as libraries for data manipulation and model evaluation. Separation of frontend UI from backend processing is clearly defined to allow one to take advantage of modularity and maintainability.

4.2 Data Collection and Preprocessing

The data collection and preprocessing phase is filled with a large dataset of Medicare patients gathered together with rich demographic, clinical, and historical health data. Techniques of preprocessing are applied to ensure data quality that include the cleaning process, handling missing values, normalizing numerical features, and encoding categorical variables.

4.3 Feature Selection

Feature selection primarily deals with the identification of features that best predict the disease. Some of the methods that can be applied in this direction include correlation analysis, recursive feature elimination, and others. Along with this, some method of dimensionality reduction is also performed so that the risk of overfitting is avoided but as much effectiveness of the model as possible is preserved.

4.4 Building the Model

The approach used for the development of the model is multi-algorithm. Support Vector Machines are used for predicting diabetes and Parkinson's disease, whereas logistic regression has been taken into consideration for predicting heart diseases. These models will learn the underlying patterns and correlations between different features and diseases by being trained with labeled input data.

4.5 Model Validation

All this is done by doing model validation against a separate dataset while testing the trained models. Metrics like accuracy, precision, recall, and F1-score are measured in an effort to find out the level of performance that the models have achieved. If the

initial performance fails to meet the required thresholds, then the action of hyperparameter tuning takes place for increasing the robustness of the model.

4.6 UI Development

The user interface will be designed as a friendly and intuitive interface using Streamlet to permit doctors to select disease types for inputting patient-specific information. In real time, validation of the user input is used in order to have reliable data, and, in case an error is observed, the error messages will pop up in real-time.

4.7 Integration and Testing with the Model

Model integration integrates the trained machine learning models into the Streamlet application to make real-time predictions based on the inputs that a user may give. Functional testing of the application will be done afterward to confirm if indeed it gives a correct prediction and an interface that is friendly to use.

4.8 Deployment and Feedback

This will be the last stage, where the application goes into live service, observed in practice with its performance, and continuously accommodates improvement by making sure that feedback from users is taken. This would identify areas for improvement, thus ensuring the predictive analytics system would remain effective and relevant to clinical practice.

5 IMPLEMENTATION PROCESS

5.1 Frontend Development Using Streamlit

The development of the predictive analytics application begins by setting up Streamlit in creating a very intuitive user interface. The front-end model is developed to enable the user to select which disease to check, such as Diabetes, Heart Disease, or Parkinson's. As per this disease selected, design a user input form to collect necessary data such as age, medical history, and any other information that may be vital for prediction.

5.2 Input Validation

The application uses input validation to ensure that the entries made by users will fall within legitimate ranges and formats, like numeric ranges, such as age, glucose levels, among others. The system also employs error handling as it gives prompt feedback, in case of entry errors or incompleteness, and directs the user to correct the entries before processing.

5.3 Selecting Models According to User Choices

Conditional logic will be used to determine which of the disease models should be utilized based on a choice selected by the user. Supervised learning techniques are available within the application in order to train models based on real-life samples of data. Once chosen, the appropriate trained machine learning model is loaded to conduct inference.

5.4 Model Inference

In the model inference, the application now configures itself to format the inputted data from the user into a form that fits the selected machine learning model. The model proceeds to run its predictions based on the formatted data. The outcome is then interpreted to include probabilities or binary classifications to show whether the user is likely to be afflicted with the disease selected.

5.5 Display Results to User

The results of the model are actually shown to the users through Streamlit's display feature. When the model gives a high probability of having a disease, the message will read as follows: "The prediction shows that you have a high chance of having diabetes from what you typed. End" Otherwise, in case the prediction would indicate a low probability, the message is also compliant with this by saying, "The prediction indicates a low percent chance of having diabetes based on your input." The application can also return confidence levels or probabilities to help make users better understand the accuracy of the prediction.

5.6 Backend Integration, Testing, and Validation

Backend integration needs to ensure that the trained machine learning models are actually well-integrated with the Streamlit interface; this can typically be done

by loading the models from saved files. Functional testing is a rather complex process to ensure that the application will execute the right work flow of user inputs and selections to produce the correct predictions. Model validation ensures that the machine learning models behave consistently with their off-line evaluations.

5.7 Code Breakdown

This code starts with import and setup, as is always necessary in a Python program. This includes `os` for interaction with the operating system and `pickle` to load saved machine learning models. The backbone of this web interface is created with `streamlit`. To add the sidebar menu, the `streamlit_option_menu` will be used.

The Streamlit app is setup as follows: st.set_page_config(title='', layout='wide', page_icon='') Here, title, layout, and icon of the app are defined in order to create uniform visual presentation. Working directory is identified using os.path.dirname (os.path.abspath (__file__)) which helps find saved model files. Pretrained models for Diabetes, Heart Disease, and Parkinson's Disease are loaded using pickle.load: preparing them for inference. A sidebar created with the use of st.sidebar enhances UX by offering clearer navigation to, for example, Diabetes Prediction and Heart Disease Prediction. Figure 1 shows the Sample code for code breakdown.

```
import os
import pickle
import streamlit as st
from streamlit_option_menu import option_menu

st.set_page_config(...)
```

Figure 1: Sample code 1.

5.8 Model Loading

The working directory is set to determine the location of the script. The `pickle` module is used to load pre-trained models for Diabetes, Heart Disease, and Parkinson's Disease from disk. Figure 2 shows Sample code for model loading.

```
working_dir = os.path.dirname(os.path.abspath(__file__))
# Load models here
```

Figure 2: Sample code 2.

5.9 Sidebar Navigation

The application has a sidebar through which the user can navigate to select the desired page for predicting different diseases. Every page is designed differently to take the required user input for the disease selected and then process it accordingly.

```
with st.sidebar:
    selected = option_menu('Multiple Disease Prediction System',
                           ['Diabetes Prediction', 'Heart Disease
                               Prediction', 'Parkinsons Prediction'],
                           menu_icon='hospital-fill',
                           icons=['activity', 'heart', 'person'],
                           default_index=0)

```

Figure 3: Sample code 3.

In summary, all parts of the application-from frontend development to data validation; models will infer the result and then display it to users-so provides a fully integrated experience of disease likelihood prediction for users on the basis of their inputs. The figure 3 shows Sample code for sidebar navigation.

6 ADVANTAGES AND DISADVANTAGES

6.1 Advantages

The key advantages of the proposed machine learning approach in predictive analytics on multiple diseases include improved early detection of disease. This enables the health provider to identify high-risk patients, followed by timely intervention and ultimately better patient outcomes. Advanced algorithms can be used, including large datasets. The system can effectively analyze large datasets and identify complex patterns that might not come into view using traditional methods. Hence, this leads to better and more accurate diagnoses as well as personalized health care solutions for individualized patients. The multi-algorithm also reduces biases

inherent in single-model evaluations, and it offers greater generalizability and robustness in predictions across different populations and healthcare settings.

6.2 Disadvantages

Although it has many advantages, the implementation of predictive analytics in the health sector presents several challenges. The first major challenge relates to the quality of data required for effective predictive analytics; low-quality data always affects prediction accuracy, which has serious implications on patient safety. Third, a problem will be the model complexity while training and the issue of feature selection, which can be challenging to make the model interpretable by healthcare providers to grasp the rationale behind the prediction. Another problem will be overfitting, as such models are typically very well trained on small, narrow datasets, and thus will compromise their generalizability to larger patient populations. Finally, the onset of these complex analytics systems will probably be difficult to adjust in working healthcare systems and will also pose logistically challenging integration work by health organizations.

7 SYSTEM ARCHITECTURE

Pre-labeled data is used for training machine learning models and supervised learning techniques. In case of a labeled dataset being available, one approaches supervised learning with the method of training a model wherein each example is tied up with a known outcome or label. For instance, in disease prediction cases, the training data can be about patient information with a disease or not. Figure 4 shows Workflow Diagram.

Training with Labeled Data: This software learns the relationships of the data itself and hence analyzes features like age and glucose level in relation to the existence or non-existence of a disease

Model Training: It could be considered as model training where the model looks for patterns and correlations between the input, including features and outcome, while scanning the labeled data.

Testing of the Model: Then, test the model against unseen data, based on performance measures that may be derived. All the training processes show how models predicts real-time results.

Making Predictions: The model, after successful train and test, can be applied to predict the chance of developing a disease based on information being provided by patients.

UI User Interface The frontend of this application is implemented using Streamlit. It will be an interactive environment where a user can be able to interact with the application. Major components include:

Streamlit Application: This is the frontend; here is where the user interacts with the app.

Sidebar Navigation: In this page, one would provide the user with an option to choose the disease they wish to predict.

Disease Prediction: Pages These would be different pages in which Diabetes, Heart Disease, and Parkinson's Disease would be predicted separately.

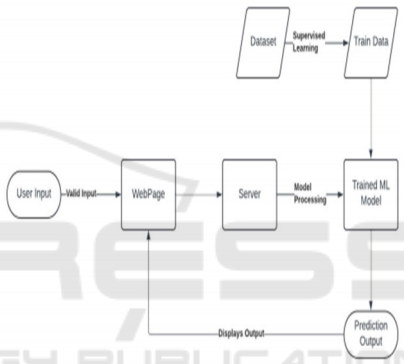

Figure 4: Workflow diagram.

Backend Components: This is where the features of backend architecture hold the loading and making of predictions via the model. Some of these include**.**

Load & Predict Diabetes Model: To load the Diabetes pre-trained model so that it will make prediction based on users' input

Load & Predict Heart Disease Model: The Heart Disease model is loaded into such a system and run with predictions

Load & Predict Parkinson's Model: Loads the Parkinson's model and makes predictions.

D. Data Flow

The application has its data flow in three stages

User Interaction: Users wish to see a type of disease predication by clicking it in the sidebar and then log data correspondingly on the specific page it leads to.

Model Processing: The data submitted for the

building of the prediction is forwarded to the concerned model of machine learning.

Results Display: On the same page of submission, the user will get the results of the prediction, displaying whether they are likely to have that disease or not.

This structured architecture clearly differentiates the user interface and backend processing, making the application better in modularity and maintainability. The figure 5 shows Use case diagram.

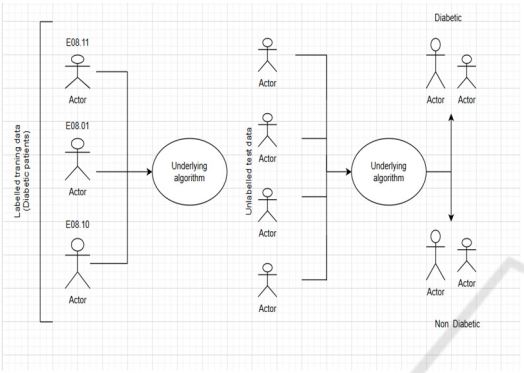

Figure 5: Use case diagram.

8 RESULTS AND CONCLUSIONS

In our multiweapon machine learning application in predictive analytics, we thus developed a multimode system that has resulted in increased accuracy in disease diagnostics with the Medicare patients. The models were trained on SVM for diabetes and for Parkinson's, while the best logistic regression emerged with the further preprocessing and feature selection of the diverse datasets. We found that, upon validation, our models generally performed very well and generalized well across demographics.

For example, in the multi-algorithm approach, it integrated not only the minimization of the inherent biases in the single-algorithm model but also comprehensive assessment of diversity in various techniques of prediction. Our system then ably managed high-quality datasets and scaled the diverse datasets so as to confront most common issues one would face when deploying a model for disease prediction: overfitting and data noise. The friendly interface set up using Streamlit has made interaction seamless and thus allows for real time prediction based on users' input.

The general aim of this project is to be able to show how predictive analytics may come to revolutionize health care through the use of chronic diseases.

Indeed, with the lessons learnt in the results from our study, such machine learning technologies shall have practical applications to offer timely and personalized solutions to patients through their respective health care providers. Work on such aspects provides initial building blocks for future innovations in the usage of machine learning within clinical applications, therefore, healthcare professionals can base their decisions on real data-driven revelations.

REFERENCES

A. Belle et al. (2015). "Big data analytics in healthcare," BioMed Research International, 2015.

A. Esteva et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks," Nature, 542(7639), 115–118.

A. Rajkomar et al. (2019). "Machine learning in medicine," New England Journal of Medicine, 380(14), 1347–1358.

A. Esteva et al. (2019). "A guide to deep learning in healthcare," Nature Medicine, 25(1), 24–29.

B. Shickel et al. (2018). "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604.

E. J. Topol (2019). "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, 25(1), 44–55.

F. Wang & A. Preininger (2019). "AI in health: state of the art, challenges, and future directions," Yearbook of Medical Informatics, 28(1), 16–26.

G. Huang et al. (2017). "Densely connected convolutional networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–47.

G. Litjens et al. (2017). "A survey on deep learning in medical image analysis," Medical Image Analysis, 42, 60–88.

H. C. Shin et al. (2016). "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," IEEE Transactions on Medical Imaging, 35(5), 1285–1296.

H. Liang et al. (2019). "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence," Nature Medicine, 25(3), 433–448.

Hodge and Austin (2020). "Applications of ensemble methods in diabetes and heart disease predictions," Journal of Medical Systems.

P. S. Kohli & S. Arora (2018). "Application of machine learning in disease prediction," 4th International Conference on Computing Communication and Automation (ICCCA), pp. 1–4.

R. Miotto et al. (2018). "Deep learning for healthcare: review, opportunities, and challenges," Briefings in Bioinformatics, 19(6), 1236–1246.
S. Uddin, A. Khan, M. E. Hossain, & M. A. Moni (2019). "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, Vol. 19, No. 1, pp. 1–16.
Sun et al. (2017). "Systematic review on preprocessing for generalization in ML," Journal of AI in Medicine.
Varma & Simon (2006). "Cross-validation for optimization," Statistical Models in Medicine.
X. Liu et al. (2019). "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," The Lancet Digital Health, 1(6), e271–e297.
Z. Zhang & L. Wang (2019). "Machine learning for clinical trials in the era of precision medicine," Contemporary Clinical Trials, 86, 105819.
Z. Obermeyer et al. (2019). "Dissecting racial bias in an algorithm used to manage the health of populations," Science, 366(6464), 447–453.