Disease Prediction Using Machine Learning Algorithms

Ivan Jovovic, Dejan Babic, Tomo Popovic, *Senior Member IEEE*, Stevan Cakic, Ivana Katnic

***Abstract*— This study aimed to investigate the application of machine learning techniques for disease prediction. Three popular machine learning algorithms, Random Forest, Support Vector Machines and Naive Bayes, were employed and their performance was evaluated. Results showed that the best performing model was based on Random Forest algorithm with the average accuracy of 87%. This model has been additionally tuned in order to achieve even better performance, which resulted with 90% accuracy. This study highlights the potential of AI in disease prediction and provides insights into the importance of algorithm selection and tuning for optimal performance.**

I. Introduction

From a mathematical perspective, learning can bedefined as gaining awareness through study, experience, analysis, or instruction. Nevertheless, when given much thought, the process of learning cannot be described by merewords, since it is a subject to change and is unique for every individual [1]. In machine learning, pattern recognition serves as the basis for computational learning. Data which is fed into it as input is used to extract knowledge or information. [2]. This field represents a crossroads between mathematics, which provides all of the methods, concepts and theories that are relevant to the field, statistics, which specializes in making predictions based on the data available, and artificial intelligence, which is nowadays a shorthand for any task which a computer is capable of performing on par with or better than an individual. With the advent of machine learning, the indivisible component of our lives has morphed into an integral part of our lives [3]. The algorithms of machine learning are strongly influencing us, whether they are used for selecting a movie or a TV series to watch or guiding you through the process of resetting your mixer (chatbots). There are a number of fields that are being impacted by this evolution, including transportation, gaming, environmental protection, security, media, healthcare, and the list is endless [4]. It is estimated that the world will have ashortage of 12.9 million health care workers by 2035,which if not addressed soon may have serious implications for the health of billions of people around the globe. Recently, we have felt the impact of these predictions on our own skin. During the Covid-19 pandemic the shortage of health-care workers and the fact the experienced doctors were overwhelmed showed us that they are in a desperate need for help. The assistance of any kind in the field of medicine, especially when there is a distressing situation, may be of a great help and could possibly help save lives of many people [5].

A number of research studies have already shown that artificial intelligence is capable of performing as well as or better than humans at a number of key healthcare tasks, including diagnosing disease. Algorithms are already outperforming radiologists in identifying malignant tumors, and in guiding researchers in how to develop cohorts for the development of expensive clinical trials [6]. As a result of the Entilic's company deep learning platform, doctors are able to gain a better understanding of a patient's real-time needs through the analysis of unstructured medical data such as radiology images, blood tests, EKGs, genomic data [7]. In an interesting development, Harvard University's teaching hospital, Beth Israel Deaconess Medical Center, used artificial intelligence to identify potentially deadly blood diseases as early as possible. The development of artificial intelligence enhanced microscopes has enabled doctors to scan for harmful bacteria such as E. coli and Staphylococcus in blood samples at a faster rate than is possible with manual scanning. A total of 25,000 images of blood samples have been used by the scientists in order to teach the machines how to search for bacteria [8]. It is evident that preventing diseases from spreading or even diagnosing them is a more effective approach than fighting them, and for this purpose, artificial intelligence has proven to be a valuable tool. By increasing the accuracy of diagnosis and thereby improving the quality of care, they have proven to be a valuable resource for the healthcare industry [9][10].

This experimental study is focused on developing a highly accurate disease prediction model using a dataset containing multiple records of documented diseases and symptoms for

*Research supported in part by the EUROCC project, European High-Performance Computing Joint Undertaking (JU) under grant agreement No. 951732. The JU receives support from the EU's Horizon 2020 research and innovation programme and EUROCC project participating institutions.

Ivan Jovovic is with Faculty for Information Systems and Technologies, University of Donja Gorica, Podgorica, Montenegro (e-mail: ivan.jovovic@udg.edu.me).

Dejan Babic is with Faculty for Information Systems and Technologies, University of Donja Gorica, Podgorica, Montenegro (e-mail: dejan.babic@udg.edu.me).

Tomo Popovićcis with Faculty for Information Systems and Technologies, University of Donja Gorica, Podgorica, Montenegro (e-mail: tomo.popovic@udg.edu.me).

Stevan Cakic is with Faculty for Information Systems and Technologies, University of Donja Gorica, Podgorica, Montenegro (e-mail: stevan.cakic@udg.edu.me).

Ivana Katnic is with Faculty of International Economics, Finances and Business, University of Donja Gorica, Podgorica, Montenegro (e-mail: ivana.katnic@udg.edu.me).

979-8-3503-9751-2/23/$31.00 ©2023 IEEE

each disease, in combination with three machine learning classification algorithms. In order to achieve as high a level of accuracy as possible, the model that achieves the best results will be tuned further. During times when doctors are overburdened, this could be an effective method of assisting them with diagnosing diseases.

II. Materials and Methods

A. The Dataset

Several young researchers from around the world have contributed to the creation of the dataset used for this research through Kaggle [12]. This dataset contains two csv files. One csv file is for model training purposes and the other one is used for evaluation of the model. The train dataset contains of 133 columns out of which 132 columns are representing symptoms and the 133th column is a disease. The dataset contained 42 diseases and 120 examples of different symptoms for each of the diseases. This in total means that this original dataset has 5040 records. So, on the first look this dataset is completely balanced, as it can be seen in Table. I and it seemed ready for the further work. With further dataset analysis it has been discovered that original dataset contains only 3431 unique records and the remaining of 1609 records are duplicates. This discoverywas a really important because of the further work that includes model building and evaluation. Using original dataset will undoubtedly result with a massive overfitting in model performance and almost every model would result with a perfect accuracy of 100%. as it can be seen in this project [13].

TABLE I - Representation of the Original and Cleaned Dataset

Disease	Org	Cleaned	Disease	Org	Cleaned
Fungal Infection	120	75	Allergy	120	82
Hepatitis C	120	80	Alc. Hepatitis	120	84
Hepatitis E	120	75	GERD	120	84
Hepatitis A	120	82	Cholestasis	120	78
Tuberculosis	120	80	Drug reaction	120	88
Cold	120	83	Peptic Ulcer	120	80
Pneumonia	120	85	AIDS	120	90
Piles	120	77	Diabetes	120	87
Heart Attack	120	75	Gastroenteritis	120	85
V. Veins	120	85	Asthma	120	87
Hypothyroidism	120	89	Hypertension	120	86
Hyperthyroidism	120	78	Migraine	120	82
Hypoglycemia	120	85	C. spondylosis	120	87
Osteoarthritis	120	76	Paralysis	120	75
Arthritis	120	79	Jaundice	120	77
Vertigo	120	78	Malaria	120	79
Acne	120	76	Chicken pox	120	87
UTI	120	83	Dengue	120	83
Psoriasis	120	89	Typhoid	120	85
Hepatis D	120	79	Impetigo	120	80
Hepatitis B	120	80	Cancer	120	76

The dataset contains records of 43 types of different categories and it was used for the model training and performance evaluation. Table I shows that our dataset was not completely balanced, since the original dataset had a number of duplicates. The differences between number of samples for each disease should not result with model underfitting, since the differences are not that significant. Every disease has between 75 and 90 samples.

B. Data Augmentation and Pre-processing

As being said in the previous part, the original dataset was not completely suitable for our research and it needed to be pre-processed. The process of non-unique data values removal was done using the Python library Pandas.

Second part was dedicated to converting object datatypes to numerical form. Our dataset indicates that our target column, prognosis, is an object datatype. This format is not appropriate for training machine learning models. Therefore, we converted the prognosis column into a numerical data type using a label encoder. By assigning a unique index number to each label, the Label Encoder converts them into numerical form [14]. A total of n labels will result in a range of 0 to n-1 digits assigned to each label.

The last part of data preparation was dividing the dataset predictors and targeted values. Targeted value in this case was column, prognosis, and it was assigned to the Y variable. The predictors variable was contained out of all the data except for the column prognosis. This data was assigned to the X variable.

C. Tools, Methods and Model Building

The models were constructed with Python programming language alongside with Sci-Kit Learn library. Pre-processing and post- processing were performed with the usage of NumPy, Pandas, Matplotlib and Seaborn libraries. All the trainings and experiments were done remotely, on a personal computer. Training for each of the models lasted about 4 minutes. As being said in the introduction part, this research was based on building 3 different models using 3 machine learning algorithms, and then after evaluating the results of these models last part is dedicated to attempt of making the best model even better.

The first model was built using Support Vector Machine algorithm. Known as SVM, Support Vector Machines are popular as one of the most used supervised learning algorithms. They are used both for classification and regression problems. This algorithm is primarily used to solve classification problems in the field of machine learning. A SVM algorithm seeks to produce the most accurate line or decision boundary that will separate n-dimensional space into categories so that we can without any problems classify new data points in the future [15]. Hyperplanes are boundaries that define best decisions. Hyperplanes are created using SVM by selecting the extreme points and vectors. Hence, SVM is referred to as the algorithm that identifies these extreme cases as support vectors. SVM offers advantages as it is suitable for cases

with high dimensionality and custom kernels can be specified. In the training process off SVM algorithm we used next parameters: C= 0.1, Kernel= rbf, Gamma = 1.
where C parameter represents a hyperparameter in SVM to control error. The smaller the C value is the smaller error the model has. In this case we used default Kernel value, rbf. RBF is an abbreviation for Radial basis function kernel an it is mostly used in these kinds of problems. Similar case as for C stands for Gamma value [16].

The second model was build using Naïve Bayes algorithm. This algorithm uses the Bayes Theorem for probabilistic machine learning. It can be used for a variety of classification tasks [17]. In mathematics, Bayes' Theorem is used to calculate the conditional probability of a given event. It has one disadvantage, however, in the event that you do not have any occurrences of a class label and an attribute value together, then your probability estimate will be based on no real data. This means that when all the probabilities are multiplied, our result will still be a zero [18]. For the purpose of our study, we employed Gaussian Nave Bayes, where continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution (Normal distribution). For the creation of this model, we used default hyperparameters, those provided when importing algorithm from Sci- Kit Learn library.

The third model was built using Random Forest algorithm. An algorithm known as random forest is definitely of the most popular algorithms in the field of classification and regression. The algorithm generates decision trees based on different data samples, and in cases of classification and regression, the majority vote is taken [19]. In this regard, the Random Forest Algorithm has the advantage of being able to handle both continuous and categorical variables, continuous as a part of regression problems and categorical variables as a part classification. Classification problems are solved more effectively using this method [20]. This algorithm is consisted of couple different parts/steps:

1. From a data set with k records, n random records are selected
2. Every sample is constructed with its own decision tree.
3. Output will be created for each decision tree
4. Depending on the classification and regression algorithms, the final output is based on majority voting or averaging.

Some of the most important features of RF algorithm:

- The diversity of trees is based on the fact that not all attributes, variables, or features are considered when making an individual tree. Each tree is unique
- The feature space is reduced because each tree does not consider all features.
- Different data and attributes are used to create each individual tree. As a result, we can utilize the CPU fully to build random forests.
- Splitting the data on train and test is not necessary when using a random forest, since 30% of the data is not visible to the decision tree
- Because the result is based on majority voting/average voting, stability is achieved.

To create a model based on Random Forest algorithm we used next parameters: n_estimators=1000, max_depth= 20, min_samples_leaf=1. Num_estimators is a hyperparameter that represents number of decision trees in the forest. In Random Forest, a tree's maximum depth is defined as its longest path between its root node and its leaf node. Upon splitting a node, min_samples_leaf defines the minimum number of samples that must be present in the leaf node [21].

As far as the best model hyperparameters tunning we used Grid Search CV. In Grid Search CV, all of the hyperparameters and their values are mixed differently and the performance is calculated for each combination and the best value for each hyperparameter is selected [22]. Due to the presence of numerous hyperparameters, this makes the processing time-consuming and expensive.

Since Random Forest showed the best performance out of two other model. Grid Search CV was performed on this algorithm and the prepared dataset. Random Forest hyperparameter tuning is believed to be one off the most resource consuming tasks, since this algorithm has a lot of hyperparameters that could be changed.

III. Results and Discussion

As stated previously, the three models have been trained using three classification algorithms. The evaluation and the results of all three models have been represented using ROC curve and classification reports of all off the models. A receiver operating characteristic curve (ROC curve) represents the performance of a classification model across a variety of classification thresholds. In this chart, two parameters are plotted: The True positive rate and the False positive rate [23].

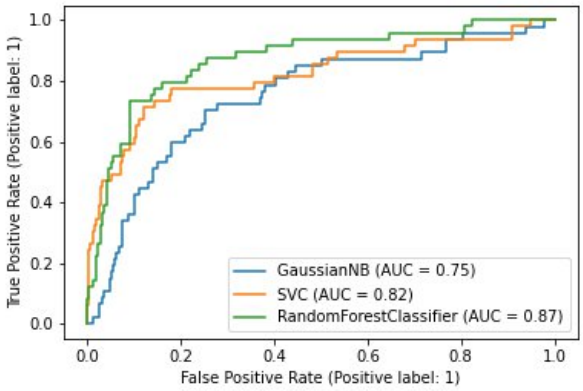

Fig. 1. ROC curves of all the models

The ROC curve is measured by AUC, which is the area in two dimensions beneath the entire curve. AUC range is between values of 0 and 1. In the case of a model that is 100% wrong in its predictions, AUC is 0.0. In the case of a model that is 100% correct in its predictions, AUC is 1. As shown

in Fig. 1. the AUC, or precision of Naïve Bayes is 75%, 82% for Support Vector Machine model and 87% for Random Forest model.

All of the results were represented in one table as it can be seen in Table II. Classification report consists 3 performance measures. Recall, F1-score and Precision. As a measure of precision, precision refers to our model'sprecision and accuracy, as well as the percentage of predicted positives that are actually positive. By labeling a true positive as Positive, recall determines how many of the Actual Positives our model captures. In order to achieve a balance between precision and recall, the F1-score must be used. By taking the harmonic mean of the precision and recall of the classifier, it is able to combine these two metrics into one.

TABLE II
CLASSIFICATION REPORT FOR THE MODELS CREATED

	Precision	Recall	F1-score
NB	0.76	0.74	0.75
SVM	0.82	0.81	0.82
RF	0.88	0.86	0.87

As we can see from the Fig. 1. and Table II. the Random Forest model showed the best performances. Next step has involved tuning the hyperparameters of the RF model and then re-training it with the new ones. Using Grid Search CV technique, we got the following parameters, represented as JSON object, that are considered best in this case:

```
{'bootstrap': True,
   'max-depth: 40,
   'max-features': 10,
   'min-samples-leaf': 5,
   'min-samples-split': 12,
   'n-estimators': 500}
```

Re-training the RF model with these parameters resulted with improvement of 3% in overall model accuracy and the final model's accuracy is a 90% as it can be seen in Fig. 2.

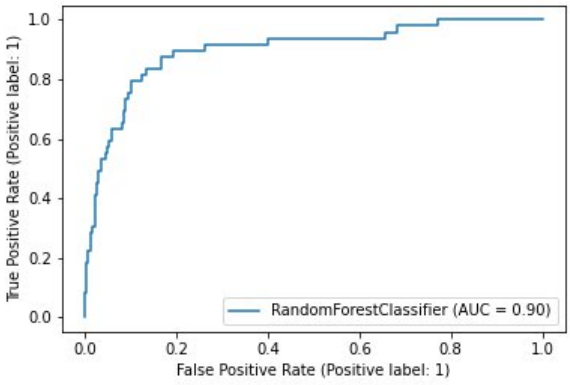

Fig. 2. ROC curve of the hyperparameters tuned model

IV. CONCLUSION

This research was focused on building the ML model that would be able to predict patient's disease based on given symptoms. It consisted part of data preparation and augmentation, building three models based on three different classification algorithms and then evaluating their results.

In our case, the model which accomplished the best results was the Random Forest model, was further tuned and trained in order to achieve better accuracy. Final model accuracy was improved by 3% using Grid Search CV technique. However, further exploring of dataset and developing of the model that uses deep neural networks will be subject of our further research in order to determine whether or not it is possible to achieve even higher accuracy then what was presented in this paper.

REFERENCES

1. W. Akinfaderin, "The Mathematics of Machine Learning", Towards Data Scince, 2017.
2. M. Haenlein, A. Kaplan, "A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence", California Management Review, 2019.
3. P. Domingos, "The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World", Basic Books, 2015.
4. A. Amisha, P. Malik, M. Pathania, V. Rathaur, "Overview of artificial intelligence in medicine", J Family Med Prim Care, 2019, pp. 2313-31.
5. World Health Organization, "Health workforce", Available online: https://www.who.int/health-topics/health-workforce, Last accessed: 30.08.2022.
6. D.Babic et al, "Detecting Pneumonia With TensorFlow and Convolutional Neural Networks", 2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS), 2022.
7. V. Kaul, S. Enslin, S. Gross, "History of Artificial intelligence in medicine", Review Article, 2020.
8. K. Minolta, "Leading AI company", [Online] Available: https://healthmanagement.org/c/it/pressrelease/leading-ai-company-enlitic-celebrates-continued-growth-with-new-products-and-partnerships, Last accessed:30.08.2022.
9. Y. Zhao et al, "Social Determinants in Machine Learning Cardiovascular Disease Prediction Models: A Systematic Review", American Journal of Preventive Medicine, 2021.
10. A. Hosny et al, "Artificial intelligence in radiology", Nature Reviews, 2018.
11. S. Daley, "40 AI in Healthcare Examples Improving the Future of Medicine", Builtin, 2022.
12. V. Danushkumar et al, "Disease Dataset", Available online: https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning.
13. V. Palkadamba, "Disease prediction using machine learning", GeeksforGeeks, 2021.
14. B. Jia, M. Zhang, "Multi-Dimensional Classification via Decomposed Label Encoding", IEEE Transactions on Knowledge and Data Engineering pp (99):1-1, 2021.
15. R. Chen, "Support Vector Machines", Artificial Inteligence, 2022.
16. D. Eyang, "Understanding Support Vector Machines", Machine Learning, 2022.
17. S. Chowdhurt et al, "Development of Naive Bays algorithm for machine learning", pp. 4-6, 2017.
18. F. Hristea, "The Naïve Bayes Model for Unsupervised Word Sense Disambiguation", Springer, 2013.
19. P. Calhoun et al, "Random Forest", Wiley StatsRef: StatisticsReference Online, 2021.
20. J. Bai et al, "Multinomial Random Forest", Pattern Recognition, 2021.
21. M.Bicego et al, "On learning Random Forests for Random Forest-clustering", 2020 25th International Conference on Pattern Recognition (ICPR), 2021.
22. G.Selvaraj, S. Brindha, "Hyperparameters Optimization using Gridsearch Cross Validation Method for machine learning models in Predicting Diabetes Mellitus Risk", 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), 2022.
23. K. Murphy, "Machine Learning: A Probabilistic Perspective", The MITPress, 2012.