Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2022, 9(3):1-6

Research Article ISSN: 2394 - 658X

Creating an Artificial Intelligence (AI) Model for Healthcare Diagnostics

Grace Camille Curtom[3], Sangeetha Madhure Nagaraju[2], Eder Mendonca[1], Mamoun Abu-Samaha[2] and Jeong Hee Kim[1]

[1]Dept. of Electrical Engineering, International Technological University, Santa Clara, California
[2]Dept. of Computer Science, International Technological University, Santa Clara, California
[3]Dept. of Engineering Management, International Technological University, Santa Clara, California
International Technological University, 3120 Scott Blvd., Santa Clara, California, 95054
Contact: +1 408 733 1623, Fax: +1 408 886 9209
Corresponding Author: Sangeetha MadhureNagaraju
Email: madhurenasange3441@students.itu.edu

ABSTRACT

The US healthcare system is a cottage industry and as it moves toward electronic healthcare data, there is no standard for healthcare interoperability, there is no proper integration among the different clinics and services leading to a lack of continuity and coordination of care with issues concerning data privacy and security. This leads to patients seeking medical services from several clinics to get the most accurate healthcare diagnosis. To solve such problems, we used artificial intelligence (AI) technology more specifically using AI modeling to obtain better and accurate diagnostics without compromising patients' data and security. As it is hard to obtain data and current AI models from the different clinics, we have obtained a public dataset from Kaggle. We have used three different available and known AI models (K Nearest Neighbor or KNN, Random Forest, and Decision Trees) to mimic three different clinics. The dataset was split into75% and 25% for training the individual AI models and testing the combined AI model respectively. These three different AI models were combined using Ensemble Technique by using a weighted average to increase their accuracy. The result of our ensemble AI model has an accuracy of 96.3% with no false negative results, helping doctors to create a more accurate diagnosis for patients with complex status. There are several advantages of creating the AI model, it is not data-dependent, and it leverages the available AI models used by the different electronic healthcare record providers; it removes the current medical bias in healthcare diagnostics; it also aims to provide a more holistic and individualized approach to healthcare diagnostics. The AI model will be able to better determine the different factors and healthcare determinants that could be more beneficial in predicting health diagnostics as well as for public health. The AI model will be able to help in the clinical decision-making inmaking more accurate healthcare diagnoses and eliminate the need for a second opinion. Thus, instead of sharing electronic health records, AI models can be shared and combined for better accuracy, helping identify the different health factors/ determinants that are beneficial in healthcare diagnostics, causing much lower healthcare costs, and aiding physicians when making clinical decisions for a more accurate diagnosis.

Key words: Artificial Intelligence Models, Healthcare Diagnostics, Electronic Health Records

---

1. INTRODUCTION

The United States Healthcare System is a big cottage industry where state-of-the-art technologies are present, but the infrastructure is insufficient. Most of the services are in disarray and they don't talk with each other. This lack of integration among the different services fails to deliver the best healthcare for the patients and the entire population in general, and also resulting in the lack of continuity and coordination of care. Also, issues regarding patient data privacy and data sharing among healthcare entities that have caused significant miscommunication and created redundant and wasteful processes. This excess processes greatly increased the costs of healthcare that resulted in increasing hospital readmission and patients' suffering. Currently, there is no single standard for healthcare interoperability, making it

difficult to transfer healthcare data from one provider to another. Patients tend to hop from one medical provider to another trying to get the same healthcare diagnostics but even with the advancement in technology with the different electronic health records, patients receive varied healthcare predictions that complicates clinical decision making and increases medical expenditure. With the advancement of information and technology primarily with the use of artificial intelligence (AI), this thesis aims to make substantial changes in improving healthcare access, quality, delivery, and service.

The purpose of this study is to determine different Artificial Intelligence (AI) models used for healthcare diagnostics and prediction and combine these models for greater diagnostic accuracy. Moreover, this will revolutionize the pathway of combining several known and available AI models in medical diagnostics to create a single AI model that will increase the accuracy of healthcare diagnostics. Lastly, help in resolving the issue of data sharing and interoperability issues for better and more accurate healthcare diagnostics that will complement medical providers in solving complex medical problems.

The importance is to show that AI modeling predictive analytics applied to clinical use can impact the future directions in care for patients especially for those that would require further assessment and validation for reproduction and generalization of results for many diseases.

2. LITERATURE REVIEW

Electronic Health Records (EHRs) are defined as digitally stored healthcare information throughout an individual's lifetime to support continuity of care, education, and research. The EHRs may include such things as; observations, laboratory tests, medical images, treatments, therapies, drugs administered, patient identifying information, legal permissions, and so on. With the growing emphasis on providing the right information to the right person anywhere at any time in today's globally interconnected world, the U.S. healthcare industry has been moving toward the EHRs system.

Several obstacles have been cited as explanations for why the EHRs have not achieved more prevalent usage in physicians' offices. These obstacles include

The EHRs products are expensive and require a major investment

The EHRs applications are not standardized

The EHRs are more difficult to use than paper-based records

The EHRs implementation reduces practice productivity and disturbs workflow (at least initially) [1].

Artificial Intelligence has been an important technology that has been adopted by many industries and is applied even to healthcare.

Artificial Intelligence is a subfieldwithin computer science associated with constructing machines that can simulate human intelligence. An AI model is a program or algorithm that utilizes a set of data that enables it to recognize certain patterns. This allows it to reach a conclusion or make a prediction when provided with sufficient information [3].

The desire to improve the efficacy and efficiency of clinical care continues to drive multiple innovations into practice, including AI. With the ever-increasing demand for health care services and the large volumes of data generated daily from parallel streams, the optimization and streamlining of clinical workflows have become increasingly critical. AI excels at recognizing complex patterns and thus offers the opportunity to transform image interpretation from a purely qualitative and subjective task to one that is quantifiable and effortlessly reproducible.

AI may also quantify information from images that are not detectable by humans and thereby complement clinical decision-making. AI also can enable the aggregation of multiple data streams into powerful integrated diagnostic systems spanning radiographic images, genomics, pathology, electronic health records, and social networks [2].

In complex machine learning problems involving multiple-base classifiers, a consensus result can be achieved using an ensemble approach which can improve the accuracy of intermediate results. This ability to combine intermediate results creates an ensemble algorithm, which has recently been the focus of ongoing research, according to various machine learning approaches applied to numerous domains employing classification, regression, and clustering [7].

Ensemble methods allow us to take a sample of Decision Trees into account, calculate which features to use or questions to ask at each split and make a final predictor based on the aggregated results of the sampled Decision Trees [4].

The ensemble techniques can be classified into two: the simple and advanced types. The simple ensemble techniques may either use max voting wherein the predictions by each model are considered as a 'vote' and the predictions which we get from the majority of the models are used as the final prediction or averaging wherein it can be used for making predictions in regression problems or while calculating probabilities for classification problems, and weighted average where all models are assigned different weights defining the importance of each model for prediction [6].

On the other hand, the advanced type of ensemble techniques may either use Bagging (or Bootstrap Aggregating) technique that uses these subsets (bags) to get a fair idea of the distribution (complete set), or Boosting wherein there is a sequential process where each subsequent model attempts to correct the errors of the previous model and the succeeding models are dependent on the previous model, or stacking is an ensemble learning technique that uses predictions from multiple models (for example., decision tree, KNN or Support Vector Machine (SVM) to build a new model, and blending that follows the same approach as stacking but uses only a holdout (validation) set from the train set to make predictions [6].

To solve this issue, the thesis will focus on combining AI models from the different EHR providers to increase the accuracy of healthcare diagnosis. This single AI model can be shared among the different EHR providers instead of sharing patients' healthcare data.

3. METHODOLOGY

Datasets

It is hard to acquire the healthcare data from different providers, hence, the Diabetes dataset from Kaggle [8] was used to realize this thesis. The data sets were clustered into different feature sets to mimic the different datasets of healthcare providers. Also, each provider was assigned a different AI model that is commonly used in healthcare diagnostics. In this thesis, the dataset we have obtained contains 8 features which are age, number of pregnancies, glucose level, blood pressure, skin thickness, insulin, BMI and diabetes pedigree function and an output which says if the patient has diabetes or not.

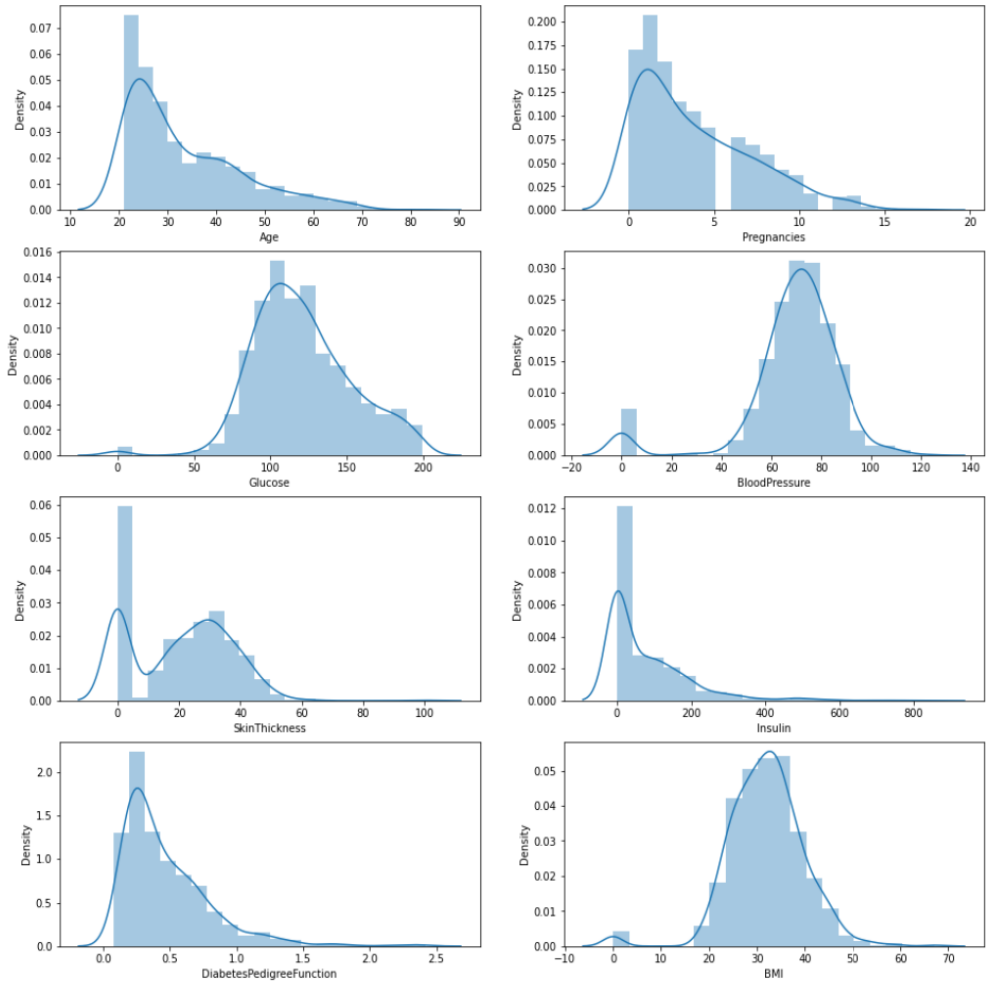

Fig. 1 Histogram of all the feature variables

Data Analysis

We try to analyze the data itself, so we get a better understanding about the dataset. We have histogram plotting for all the features, finding correlation between all the features and the diabetes output, looking for any missing values in the dataset and replacing those by their mean values, and also outlier detection and removal of such data points. We have included the histogram plots of all the feature variables in the figure 1 and the correction between features and output in the figure 2.

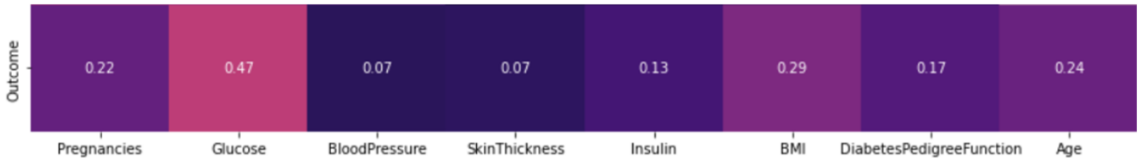

Fig. 2 Heatmap showing correlation between features and output

Feature Engineering

New variables were created according to BMI, insulin and glucose variables. NewBMI is the categorical variable created with 5 categories – underweight, normal, overweight, obesity 1 and obesity 2. NewInsulinScore is created with 2 categories – normal and abnormal. Similarly, NewGlucose is created with 5 categories – low, normal, overweight, secret and high.

One-Hot Encoding

Categorical variables in the dataset are converted to numerical data using transformations like one-hot encoding and label encoding.

AI modeling

This step consists of creating 3 individual AI models that mimics 3 different clinics

The first model that we used to mimic clinic 1 is K-Nearest Neighbor (KNN). This has the following features- Pregnancies, Glucose, blood pressure, skin thickness, Insulin, BMI, and Age.

The second model that we used to mimic clinic 2 is Random Forest (RF). This has the following features- Diabetes Pedigree Function, Glucose, blood pressure, skin thickness, Insulin, BMI, and Age.

The third model that we used to mimic clinic 3 is the Decision Tree. This has the following features- Diabetes Pedigree Function, Pregnancies, Glucose, blood pressure, Insulin, BMI, and Age.

The ensemble model is created using the weighted average method of the 3 clinic models by having a multiplication factor of the model's accuracy as its weights.

We have split the dataset into 4 sets. We used the first three sets to train the 3 models (25% each - 75% total), and the last set of data (25%) to test all the models including the ensemble model.

Assessments and Measures

We have run and checked the accuracy for each healthcare provider or clinic and identified its set of features. Important features were identified using Heatmapthat shows the correlation between each feature and the output. Finally, we have used the ensemble technique using weighted average method with accuracy of the model as its weights to aggregate models from each clinic and provide improved accuracy.

```
y_comb = []
for i in range(y[X_testing.index].size):
  w_mean = (y_pred1[i]*clinic1_acc + y_pred2[i]*clinic2_acc + y_pred3[i]*clinic3_acc)/(clinic1_acc+clinic2_acc+clinic3_acc)
  y_comb.append(1 if w_mean>=0.5 else 0)
y_comb = np.array(y_comb)
combined_acc = metrics.accuracy_score(y[X_testing.index], y_comb)
print("Accuracy = {0:.3f}".format(combined_acc))
```

```
Accuracy = 0.963
```

Fig. 3 Creating an ensemble model

4. RESULTS

The first model that we used to mimic clinic 1 with the model K-nearest neighbor (KNN) gave an output accuracy of 90%. The second model that we used to mimic clinic 2 with Random Forest (RF) model gave an output accuracy of 94.2%. The third model that we used to mimic clinic 3 is the Decision Tree model that gave an output accuracy of 91.1%. The ensemble model with a weighted average of the three previous models has an improved accuracy of 96.3%

```
[ ]
    print("Clinic 1: Accuracy = {0:.3f}".format(clinic1_acc))
    print("Clinic 2: Accuracy = {0:.3f}".format(clinic2_acc))
    print("Clinic 3: Accuracy = {0:.3f}".format(clinic3_acc))
    print("Combined: Accuracy = {0:.3f}".format(combined_acc))
```

```
Clinic 1: Accuracy = 0.900
Clinic 2: Accuracy = 0.942
Clinic 3: Accuracy = 0.911
Combined: Accuracy = 0.963
```

Fig. 4 Accuracies of all three clinic models and ensemble model

We have plotted the confusion matrix for the ensemble model that shows the summary of the prediction against True Positives, False Positives, True Negatives and False Negatives. We can observe that there are no false negatives and there are 7 false positives in the prediction.

Fig. 5 Confusion matrix

5. DISCUSSION

The findings support that even with different available AI models we can combine them into one AI model that can be shared among different healthcare providers and clinics without the issue of data security and privacy.

Patients no longer need to get subsequent opinions to verify healthcare diagnostics. If two providers/clinics have already partnered with the AI model, there is a guarantee that the results will be the same for both providers.

Also, the AI model better assists in clinical decision-making among doctors. For instance, investigations leveraging computer-aided diagnostics have shown excellent accuracy, sensitivity, and specificity for the detection of small radiographic abnormalities, with the potential to improve public health. AI often detects minor image alterations, more relevant outcome variables including a new diagnosis of advanced disease, a disease requiring treatment, or conditions likely to affect long-term survival. The occurrence of clinically meaningful events—symptoms, need for disease-modifying therapy, and diseases that lead to mortality that greatly affect the quality of life is the focus of AI-based investigations [5].

Lastly, the thesis aims to provide a more holistic and individualized approach to healthcare diagnostics. Using the AI Ensemble technique, medical, dental, and ophthalmological health records can now be combined to help doctors for better decision making and more accurate healthcare diagnostics.

Advantages

There are several advantages to using AI models. Some of them are as follows: (1) The AI model is not data dependent. The AI model leverages in assessing all the different factors and determinants available from the different electronic healthcare record providers. (2) There is no need to create new models. The AI model leverages the existing AI models of the different healthcare providers and clinics. (3) The AI model removes the current medical bias in healthcare diagnostics. The AI model is aimed to create a framework that is unbiased by developing and testing rigorously so as not to create systemic injustice in healthcare prediction. (4) The AI model aims to provide a more holistic and individualized approach to healthcare diagnostics. As we are not data-dependent, healthcare records from different sources can be combined to create a more accurate healthcare diagnostics and outcome prediction for patients. (5) The accuracy of both the individual AI models and the combined AI model will improve with more AI models added. (6) The AI model will be able to better determine the different factors and healthcare determinants that could be more beneficial in predicting health diagnostics as well as for public health. (7) The AI model will be able to help in the clinical decision-making when making healthcare diagnosis. This aids in training and testing doctors. (8) Eliminating the need for a second opinion, reducing redundancy for further testing and treatment can lower the overall cost of healthcare.

6. CONCLUSION

This thesis paper is aimed to identify the different AI models used for predictive healthcare diagnostics. Based on the research, the different AI models can be shared among the different healthcare entities solving the patients' data security and privacy issues. AI models can be combined using the Ensembling Technique, which optimizes and increases the accuracy of the different AI models. The AI model can help identify different health factors/ determinants that are beneficial in healthcare diagnostics. This lowers the cost of healthcare and aids in physicians' clinical decision-making in diagnosis.

7. RECOMMENDATION

It is recommended that we can create our own AI model as we acquire expertise from different models and results and fine-tune the model to improve the overall accuracy. It is also recommended to leverage the same business model to different industries where AI is used by different competitors and provide increased accuracy from the combination, benefiting the whole industry outcomes. Furthermore, it is to try to increase the synergy between different areas from the

same industry, e.g., dentist and gastric clinic, as the features from one can influence the outcomes of the other and can help to achieve better accuracy.

Acknowledgement
We would like to express our sincere gratitude to our beloved Professor, Mamoun Abu-Samaha, International Technological University, Santa Clara for providing us with this opportunity to carry out this thesis work. We are deeply indebted to our professor for his continuous guidance, motivation and valuable support throughout our coursework which helped us in completing this thesis successfully. We would like to also express our sincere gratitude to Professor John Kim, International Technological University, Santa Clara who helped us in our thesis through his guidance, feedback, and reviews. We would also like to express our heartfelt thanks to our team members for their involvement and contributions throughout this thesis work. We would like to thank our classmates who were supportive and helped us in our thesis work.

REFERENCES

[1]. S Ajami S & R Arab-Chadegani R, Barriers to implement Electronic Health Records (EHRs).*Mater Sociomed*, 2013, 25(3):213-5.
[2]. WL Bi, A Hosny, MB Schabath…&... H Aerts, Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: a cancer journal for clinicians*, 2019, 69(2), 127–157.
[3]. N Klingler, What is an AI model? Here's what you need to know, V*iso.ai.* https://viso.ai/deep-learning/ml-ai-models/, 2021.
[4]. E Lutins, Ensemble Methods in Machine Learning: What are they and why use them?, *Medium.* https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f., 2017.
[5]. O Oren, BJ Gersh, & DL Bhatt, Artificial Intelligence in Medical Imaging: Switching from Radiographic Pathological Data to Clinically Meaningful Endpoints. *The Lancet Digital Health*, 2020, *2*(9).
[6]. A Singh, Ensemble learning, Ensemble techniques, *Analytics Vidhya*, Web. https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/, 2018.
[7]. C Song, A Pons, & K Yen, Sieve: An Ensemble Algorithm Using Global Consensus for Binary Classification. *AI*. 1. 242-262. 10.3390/ai1020016, 2020.
[8]. UCI Machine Learning, Pima Indians Diabetes Database, *Kaggle*, Web. https://www.kaggle.com/uciml/pima-indians-diabetes-database, 2016.