Cancer Detection and Classification using Random Forest, NN and XGBoost Algorithm of Machine Learning

Manasa T P [1*††] **and, Dr. Mohammed Tajuddin** [2 *†]
Department of Computer Science & Engineering
**Dayananda Sagar College of Engineering*
[†]*Bangalore Institute of Technology*
[†]Visveswaraya Technological University, Bengaluru, Karnataka, India
tpmanasa@bit-bangalore.edu.in
tajuddin-cs@dayanandasagar.edu

Abstract— Healthcare is the topmost high priority subject matter spoken about. Detailed diagnosis of the disease is required to give individuals the required attention in time. Advancements in healthcare using ML technologies have been beneficial for the health component of exploration and early diagnosis. The extremely high number of new methodologies permits new pathways favorable for the society. The foremost period is the most crucial which when detected early can be easily curable. Branches of machine learning which include deep learning whose applications have lately gained more importance in medical text and image research due to their benefits and success. Step has been taken to assess studies on ML and DL methods used to mimic different types of cancer using these three categories including the intent of the prediction, approach of prediction, and data instances. For this, we suggest systematic examination of different human cancerous diseases by applying techniques such as NN, XGBoost, Random Forest to make important predictions and help in decision-making. By using this methodology, the system proposed in this work is able to achieve 99.45% accuracy and 99.95% AUC in detecting whether the patient has cancer or not and achieved 93.94% accuracy in classifying the cancer types. Our methodology has been successful in solving the patients problem with recommendable results.

Keywords— **Human Healthcare, Random Forest, Recognition, CNN, XGBOOST**

I. INTRODUCTION

Machine learning (ML) has triggered broader fields in the electronic space of handheld devices to industry machineries, definitely in education removing the dependence for students, healthcare regions fostering early detection and escalating recovery. Researchers have demonstrated how machine-learning-based disease diagnosis (MLBDD) can be both charged effectively in less time. Conventional pattern of diagnosis is much lagging of time , overpriced and always needs human association. While human limitations decays the procedure, ML-based systems have been recognized as upper hand comparatively, as machines work effortlessly. To create MLBDD setup, pixelated healthcare data mostly containing X-rays and MRIs whereas tabular data storing patient conditions, ages, and genders are used. Artificial intelligence can help by providing specialized individual medication to escalate patient treatment. AI methods, including machine learning wherein deep learning is more advanced, are widely applied for healing, disease judgement, pill discovery, and patient risk detection. Full disease diagnosis using AI techniques like ultrasound, magnetic resonance imaging, mammography requires various medical data sources. Furthermore, AI has significantly improved the clinic experience and accelerated patients' readiness for rehabilitation treatment at home. This paper presents a detailed study on technology which uses computer intelligence for verification and evaluating various cancer diseases. Results are set side by side to find the contrast using different quality measures like prediction rate and precision. Introduce systematic machine learning executable strategy containing neural networks, XGBoost, and Random Forest to make significant predictions and assist in decision-making.

Increase in cancer cases has been mapped to various elements, including unhealthy diets, obesity, genetics [1]. In our bodies, cells have an organized structure of splitting and growing. These patterns are occupied within deoxyribonucleic acid configured as genes. However, when mutations occur, cells can grow uncontrollably, leading to malignant tumors, or cancer. Colorectal cancer possesses considerable health menace [2,3] and Saudi Arabia is one associated with this condition [4]. About 90% of colorectal cancer cases are found in people over 45 years old [5], and men have the highest rates of this cancer [6]. Colorectal cancer is allied to deadliest malignant tumors, causing the deaths of over 4,000 people each year in the Kingdom of Saudi Arabia [7].

Following are the goals demanded to be achieved for classification:

• Patterns from the user data are recognized and captured and used for predictions and classification tasks.

• To effectively predict if the patient suffers from the disease like various cancerous diseases.

• To make early decisions which will ensure the patients are attended with perfectly designed medication specific to their diagnosis.

Artificial intelligence has improved life in the healthcare field. This improvement has raised both the quality and efficiency of systems and services related to health [8]. AI can serve as a solution for predicting colorectal cancer by leveraging intelligent data driven models. Beforehand detection of such diagnosis with great accuracy is achievable using ML techniques. This leads to a higher success rate for treatment and lowers the colorectal cancer mortality rate [9]. This approach also enhances the chances for thriving a high standard of life for colorectal cancer patients [10].

II. LITERATURE REVIEW

MIN CHEN et al. [1] proposed a disease prediction system in their paper was able to get an accuracy of 94.8% which was achievable with the help of multiple machine learning algorithms.Sayali Ambekar et al. [2] recommended cancer threat vaticination and used a convolutional neural network. Various ML techniques have been successfully flourished by adopting CNN-UDRP algorithm, KNN algorithm and Naive Bayes. 82% accuracy evaluation has been achieved by using Naive Bayes method. Naganna Chetty et al. [3] developed a system that provided better results for disease prediction by using a fuzzy approach. They applied techniques like the KNN classifier, Fuzzy c-means clustering, and Fuzzy KNN classifier. Their paper focused on predicting diabetes and liver disorders, with an accuracy of 97.02% for diabetes and 96.13% for liver disorders.

Generally, unsupervised knowledge aims to find sheltered patterns in data and use them to infer rules. Unsupervised knowledge carrying unlabeled values which is why patterns have to be recognized which becomes a major challenge. One well- known illustration of an unsupervised algorithm is k- means clustering. Semi- supervised knowledge offers a way to combine the advantages of supervised and unsupervised knowledge. In the first two orders of affairs, labels are also handed for every observation or none at all. Sometimes, certain obediences admit labels, but utmost are left unlabeled because labeling costs capitalist and needs specialized knowledge. Semi-supervised algorithms are swish for creating models in these situations. type, regression, and prophecy problems can all be addressed with semi- supervised knowledge.

Data Collection:
For the purposes of this study, we employed a blood- predicated dataset first presented by Cohen et al., which contains information on DNA mutations, protein biomarker cornucopia, and a range of clinical characteristics. The dataset consists of 1817 blood test samples, of which 1005 belong to cancer- diagnosed cases, with an average age of opinion of 63 times. Cancers included in the dataset are colorectal, bone, ovarian, lung, liver, pancreatic, stomach, and esophageal cancers chosen because said cancers are n't generally diagnosed in an early stage through traditional blood test styles. Added to the dataset are 812 blood examples holding healthy subjects with an average age of 49 years. These individualities had no former history of cancer, severe dysplasia, autoimmune conditions, or habitual order cancer. Importantly, none of the cases had endured chemotherapy before furnishing blood samples, and no cases involved visible distant metastases at the study's onset. The arrangement of samples is displayed showing the counts for each type and healthy controls is illustrated in Figure 1.

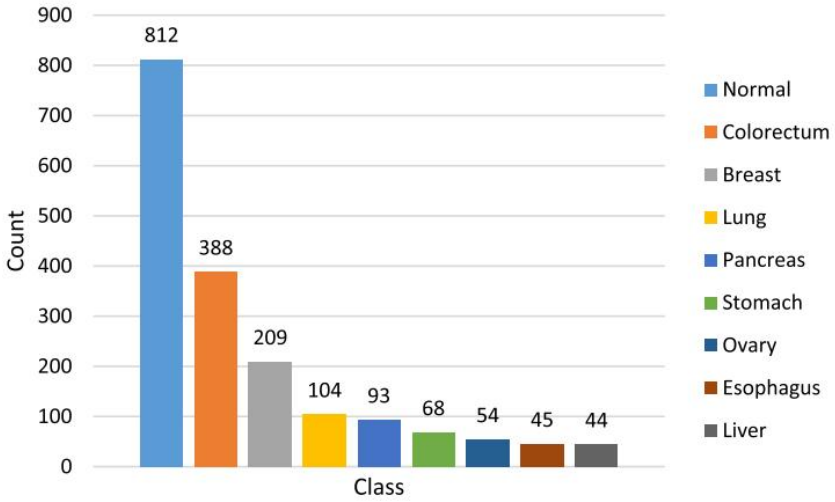

Figure 1. Dataset classes count.

Dhiraj Dahiwade et al. [4] created a model to predict cancers using machine knowledge styles. They employed ways like KNN and CNN. Case's symptoms are acquired for prognosis of the disease. The delicacy of KNN has a 95%, while CNN achieves 98%.

Lambodar Jena et al. [5] worked on predicting the trouble of habitual conditions by employing ML classifiers. They used methods analogous as Naive Bayes and Multilayer Perceptron. Their paper aims to predict habitual order cancer, with Naive Bayes and Multilayer Perceptron achieving rigor of 95 and 99.7, singly.

Dhomse Kanchan B. et al. [6] examined special cancer prophecies through top element analysis and machine knowledge algorithms. They used ways like Naive Bayes type, Decision Tree, and Support Vector Machine. The delicacy of their system is

34.89 for diabetes and 53 for heart cancers.

III. BLOCK DIAGRAM

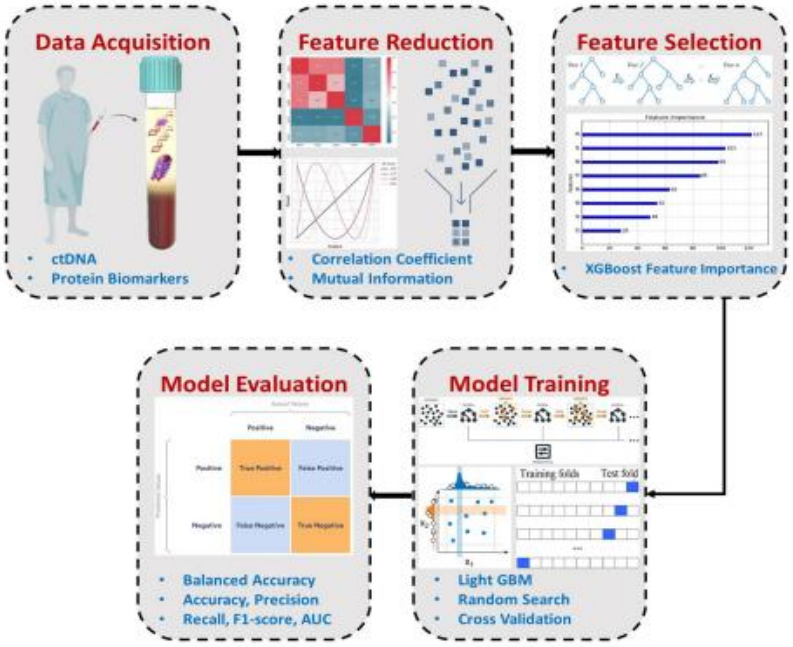

Figure 2. Precision cancer classification Model

1. Input: We are taking input from the user of the symptoms of their current cancer for the further training process.

2. Get Data: Now the user lays down the symptoms and this will be fed as the data.

3. Data Acquisition: Data collection and processing that performs two operations. The first collects the data, the alternate processes the data and extracts applicable information depending on the collected data.

4. Get Symptoms of the Body: Collecting and analyzing physical symptoms. This information is acted on by an algorithm to help predict possible illnesses.This panel collects physical function problems related to symptoms. This is analyzed to get the possible conditions.

5. Dataset cancer(symptoms, functions): In this section has a predefined cancer data set containing symptoms and features caused by conditions. This record is further used to match data entered from the user, and if there is a correct match the system suggests a possible illness.

6. Train Data(): Our cancer prophecy system is trained with the backing of various classifiers including Random Forest Classifier and XGBoost algorithm to break problems which will be suitable to distinguish between the different types of cancers in the dataset.

7. Prev_Disease(Dataset cancer): A new record from the dataset is passed as a parameter which acts as a base for predicting the further training process.

8. prophecy (symptoms, function): For this field, prophecy is performed using an arbitrary timber classifier, the XG Boost algorithm, Symptoms and their functions in the user's body are involved in prognostications.

9. Possible cancer(symptoms, function): In this field the symptoms and features are passed as parameters and the possible conditions are calculated predicated on these parameters.

10. Data Processing: This field includes the five data recycling fields and is a pivotal part of our cancer prophecy system. It has all the fields you need to exercise your data.

11. After Data Acquisition and Processing: Returns results of that type of cancer it is predicted.

IV. IMPLEMENTATION SYSTEM

Organized approach is designed to detect various cancerous diseases like breast cancer and other organ cancers. Various

datasets are pulled from Kaggle's machine learning database to implement the disease detection system. The classification computation is done with a multi classifier, of which random forest is applied for disease detection systems. Algorithms generated have led to epidemic identification in disease detection with maximum accuracy, precision and recall.

V. ALGORITHMS

1] XGBOOST:

XGBoost algorithm involves a decision tree based technique that contains a gradual increasing learning framework. In prognosis situations where unstructured values are present, artificial neural networks tend to outperform all other algorithms or frameworks. However, when it arrives at structured values, the decision tree can be measured as the best-in-class right now.

2] Random Forest:

Random Forest Classifier is a broad algorithm for machine learning, building forests of decisions and creating trees to perform classification. For arriving at a single decision it intensifies more trees each step while training and votes for higher accuracy and reduced errors. Each tree in the forest is trained with a random subgroup of data records from the original data scanned during exchanges, a method known as boat traps. Second, only randomly selected subgroups of traits are used for trees in division, introducing diversity of individual trees, preventing them from being sharply correlated. As soon as all the trees are trained, the classifier makes predictions by voting for all the trees as a result, and the most selected selection becomes the resulting prediction. This collective decision-making process makes random forests extremely reliable, especially with loud data, reducing the likelihood of over adaptation compared to trees that create a single decision.

3] Convolutional Neural Networks:

This network belongs to a deep learning strategy that can look at an image, when carefully calculated with different parts of it like learnable weights and biases, which can tell the difference between them. Layers initially contain stats on how to modify the data to get the optimized results reducing the complexity for the user and also lowering preprocessing required than other classification algorithms. In primitive methods, values of weights assigned by hand require enough training, CNN layers can learn these filters, weights and characteristics. It can be perceived similarly to how neurons function in the human brain.

VI. RESULT

1 Cancer Detection and Classification:

Three different machine learning models were trained and evaluated on the dataset to compare their performance.

Random Forest Model:

Random Forest Model Accuracy: 0.7939560439560439

	precision	recall	f1-score	support
Breast	0.68	0.64	0.66	42
Colorectum	0.69	0.87	0.77	78
Esophagus	1.00	0.11	0.20	9
Liver	1.00	0.11	0.20	9
Lung	0.53	0.38	0.44	21
Normal	0.89	1.00	0.94	163
Ovary	1.00	0.73	0.84	11
Pancreas	0.81	0.72	0.76	18
Stomach	0.00	0.00	0.00	13
accuracy			0.79	364
macro avg	0.73	0.51	0.54	364
weighted avg	0.77	0.79	0.76	364

Figure 3 : Random Forest Model Accuracy of all cancerous disease

XGBoost Model Accuracy:

XGBoost Model Accuracy: 0.8598901098901099

	precision	recall	f1-score	support
Breast	0.70	0.83	0.76	42
Colorectum	0.79	0.87	0.83	78
Esophagus	1.00	0.33	0.50	9
Liver	0.83	0.56	0.67	9
Lung	0.56	0.48	0.51	21
Normal	0.98	1.00	0.99	163
Ovary	1.00	0.82	0.90	11
Pancreas	0.81	0.94	0.87	18
Stomach	0.75	0.23	0.35	13
accuracy			0.86	364
macro avg	0.82	0.67	0.71	364
weighted avg	0.86	0.86	0.85	364

Figure 4: XGBoost Model Accuracy of all cancerous disease

Neural Network Accuracy:

Neural Network Accuracy: 0.7857142857142857

	precision	recall	f1-score	support
Breast	0.66	0.79	0.72	42
Colorectum	0.78	0.68	0.73	78
Esophagus	0.40	0.44	0.42	9
Liver	0.83	0.56	0.67	9
Lung	0.46	0.52	0.49	21
Normal	0.93	0.94	0.93	163
Ovary	1.00	0.82	0.90	11
Pancreas	0.83	0.83	0.83	18
Stomach	0.21	0.23	0.22	13
accuracy			0.79	364
macro avg	0.68	0.65	0.66	364
weighted avg	0.79	0.79	0.79	364

Figure 5 : Neural Network Accuracy of all cancerous disease

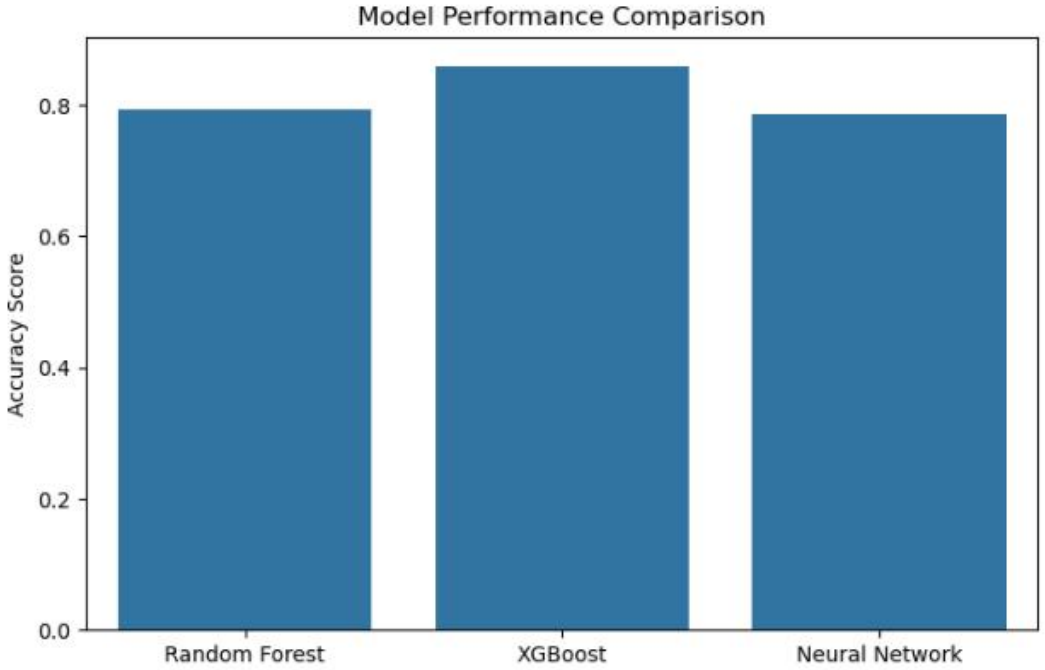

Figure 6 : Comparison of Random forest, XG Boost and Neural Network with Accuracy Parameter

```
Test Sample Index: 244
Predicted Cancer Type: Colorectum
Actual Cancer Type: Colorectum
Sample Features: [8.0705400e+02 2.1309000e+03 1.4961900e+03 7.3600000e+00 3.4300000e+00
 1.6380000e+01 1.1750000e+01 5.7199000e+03 2.1069700e+03 2.2100000e+00
 1.3472000e+03 2.8753000e+02 7.8907000e+02 5.2950000e+01 3.9006000e+02
 1.9400000e+00 6.8279400e+03 2.2931100e+03 7.7300000e+01 6.0540000e+01
 3.9246700e+03 3.5103190e+04 1.6740000e+01 6.9729000e+02 1.2878000e+02
 4.3560000e+01 4.4000000e-01 1.5449431e+05 1.0375520e+04 1.1317090e+04
 1.3755100e+03 1.4509600e+03 3.8480000e+01 4.7236200e+03 5.7855800e+03
 6.6390000e+01 9.0659000e+02 8.8779640e+04 4.2531090e+04 1.5700000e+00
 3.5000000e+01 3.0000000e+00 0.0000000e+00]
```

Figure 7: Test samples details with Predicted cancerous disease as Colorectum

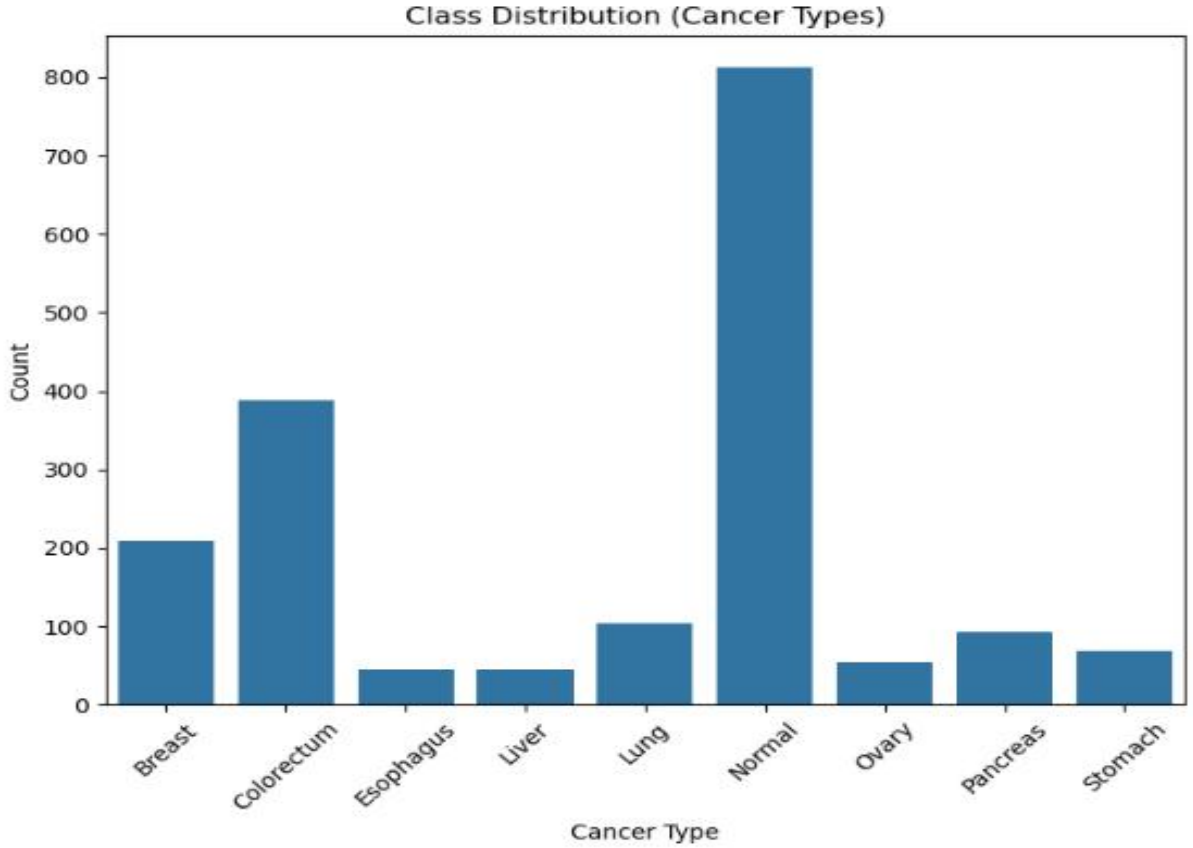

Figure 8: Cancerous distribution count with cancerous type

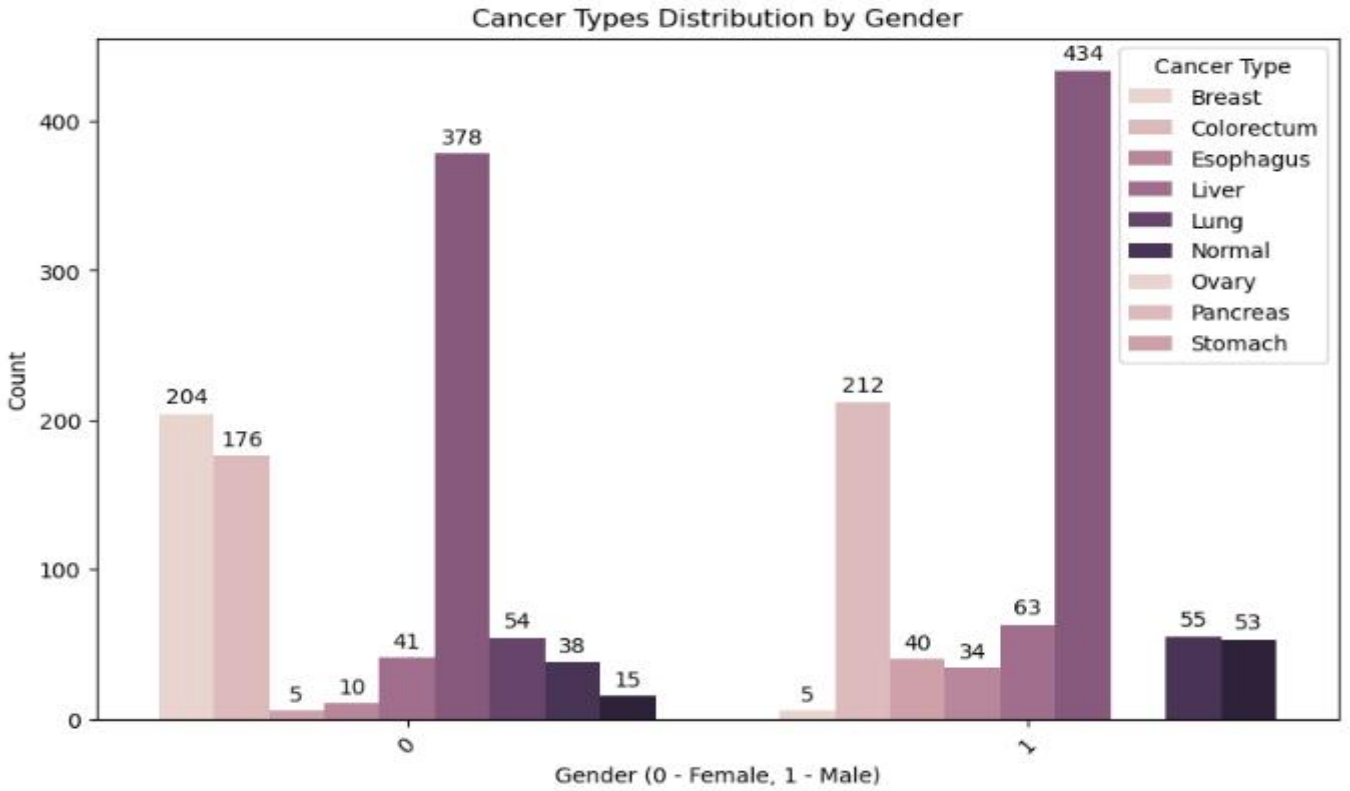

Figure 9: Cancerous types distribution by Gender

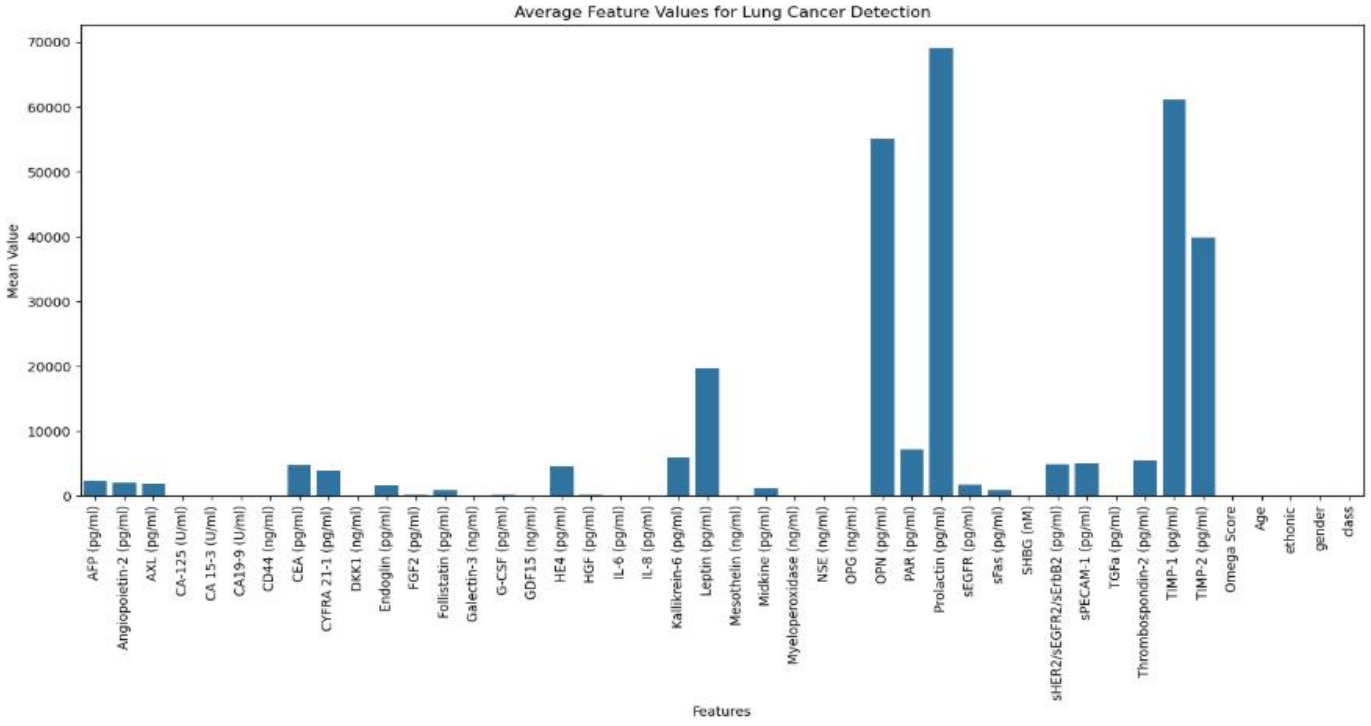

Figure 10: Average feature values for Lung Cancer Detection

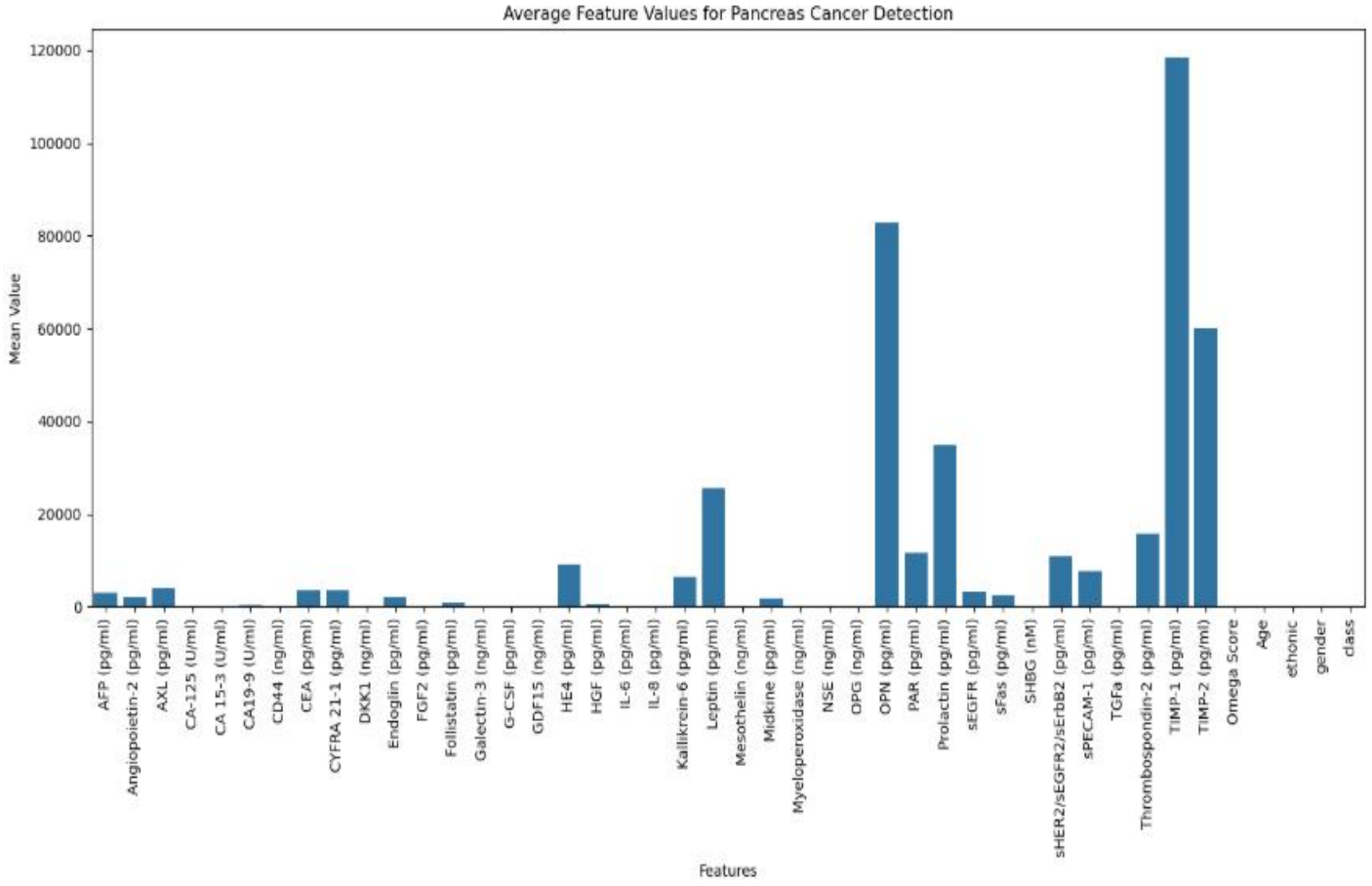

Figure 11: Average feature values for Pancreas Cancer Detection

VII. CONCLUSION

Based on our different algorithm implementations, we found NN, Random Forest, and XGBoost are convenient for evaluating diseases using the provided data set. We achieved accuracies of 73.54% for Cancerous Diseases, 94.15% for Breast Cancer, and 95% for Brain Cancer. Given our objective, we can say that we met all goals, as we predicted diseases based on input symptoms using various algorithms. Reviewed recent papers on the capabilities of ML and DL for detecting and diagnosing cancer. To make the review clearer, steps have been taken to diagnose different types of cancer using these three categories including the intent of the prediction, approach of prediction, and data instances. In each category, we provided summaries of the studies from different viewpoints. We analysed from the recorded studies focused on building predictive models using ML or DL methods aimed at predicting either a normal or abnormal state with either public or collected data sets. Finally emphasizing on curative factors which are used for examining the challenges and opportunities in utilizing ML and DL for predicting various cancer. In conclusion, AI significantly impacts healthcare by improving the prediction of various cancers through ML and DL algorithms.

VIII. REFERENCES

[1] R. L. Siegel, K. D. Miller, A. Jemal, "A Cancer Statistics" J. Clin. 67,7–30, 2017.

[2] Liu ZA, Hanley JA, Saarela O, Dendukuri N, "A conditional approach to measure mortality reductions due to cancer screening: measuring mortality reductions due to screening", Int Stat Rev 2015.

[3] Fabio Pittella-Silva, Yoon Ming Chin, "Plasma or Serum: Which Is Preferable for Mutation Detection in Liquid Biopsy?", Clinical Chemistry, Volume 66, Issue 7, July 2020.

[4] William D. Hazelton and E. Georg Luebeck , "Biomarker-Based Early Cancer Detection: Is It Achievable?", Anal. Chem. 2021.

[5] Alix-Panabieres C, Pantel, K., "Circulating tumor cells: Liquid biopsy of cancer. Clin. Chem. 2013, 59, 110–118.

[6] Chiu, T.K.; Chou, W.P.Huang, S.B., Wang, "Application of optically-induced-dielectrophoresis in microfluidic system for purification of circulating tumor cells for gene expression analysis—Cancer cell line model". Sci. Rep. 2016.

[7] National Comprehensive Cancer Network. Breast Cancer Screening and Diagnosis (Version 1.2019). Published 2019.

[8] National Comprehensive Cancer Network. Colorectal Cancer Screening (Version 2.2020). Published 2020. Accessed July 29, 2020.

[9] National Comprehensive Cancer Network. Lung Cancer Screening, Published 2020. Accessed July 29, 2020.

[10] Saslow D, Solomon D, Lawson HW, et al; American Cancer Society; American Society for Colposcopy and Cervical Pathology; Am J Clin Pathol. 2012.

[11] Borrebaeck, C. A. (2017). Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer. Nature Reviews Cancer, 17(3), 1999

[12] Joshua D. Cohen, Ammar A. Javed , Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers PNAS 114 (38) 10202-10207,2017;

[13] Kim, J., Bamlet, W. R., Oberg, A. L., Chaffee, K. G., Donahue, G., Cao, X. J., ... & Zaret, K. S. (2017). Detection of early pancreatic ductal adenocarcinoma with thrombospondin-2 and CA19-9 blood markers. Science translational medicine, 9(398). BMB reports 2008.

[14] Jiang, Y. M., Yu, D. L., Hou, G. X., Jiang, J. L., Zhou, Q., & Xu, X. F. Serum thrombospondin-2 is a candidate diagnosis biomarker for early non-small-cell lung cancer. Bioscience reports, 39(7), BSR20190476,2019.