Exploring Cancer Risk Patterns with EDA and Predictive Machine Learning Models

[1]Md Reduanur Rahman, [2]Sufia Zareen, [3]Nasrin Akter Tohfa, [4]Md Abdul Alim, [5]Md Habibul Arif, [6]Iftekhar Rasul, [7]Mamunur Rahman, [8]Md Shakhawat Hossen

[1]Master's in Information Technology, [2]Master's in Information Technology and Management, [3]Master's in Information Systems Security, [4]Master's in Information Technology in Management, [5] Master's in Information Technology, [6]Master's in Information Technology Management, [7]Master's of Science in Information Technology, [8]Master's in Information Technology

[1,5,7,8] Washington University of Science and Technology, Alexandria, Virginia, USA
[2]Campbellsville University, Campbellsville, Kentucky, USA
[3]University of the Cumberlands, Williamsburg, Kentucky, USA
[4,6] St. Francis College, Brooklyn, New York, USA

[1]mdrahman.student@wust.edu, [2]szare476@students.campbellsville.edu, [3]ntohfa15051@ucumberlands.edu, [4]malim@sfc.edu, [5] habibularif2233@gmail.com, [6]irasul@sfc.edu, [7]mamrahman.student@wust.edu, [8]mhossen.student@wust.edu

***Abstract*—Early and accurate cancer detection still represents one of the most challenging objectives in medical diagnosis. This work introduces an advanced machine learning approach created and trained to identify potential cancer outcomes using a set of clinical and lifestyle parameters. The dataset originating from the anonymized information about real patients includes data on age, gender, body mass index (BMI), smoking or non-smoking status, duration of symptoms, and other attributes. Multiple classification algorithms, such as Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, and Support Vector Classifier, were created and developed after the transformation of the dataset and the exploratory analysis. The performance was assessed using the accuracy, precision, recall, F1-score, and AUC from ROC curve analysis. The experiment proves that SVC and Logistic Regression could perform classification perfectly, while Random Forest and Gradient Boosting also ensured outstanding predictive properties. Confusion matrices showed how the power of the ensemble and the kernel-based methods-based algorithms would lie in the minimum number of errors. Overall, the accuracy of these methods is more than 96%, which confirms the potential of such decision-support tools in promoting the diagnostic reliability for further cancer detection or integration for intelligent healthcare systems that promote better decision-making.**

***Index Terms*— Cyber Security, Transfer Learning, Vision Transformer, VGG19.**

I. INTRODUCTION

Cancer is one of the most formidable public health problems of our time; it remains a primary cause of death and claims one in six lives globally, according to the World Health Organization. Despite considerable medical innovation and progress, early diagnosis continues to be the most effective tool for reducing cancer-associated mortality. Hence, while standard diagnostic tools, such as biopsy, histopathological examination, and imaging, are functional, they are typically expensive and inaccessible, and time-consuming to test. For this reason, a large proportion of patients are diagnosed with cancer in advanced stages, which significantly impacts prognosis. Advancements in machine learning and artificial intelligence have transformed the current medical landscape by allowing automated and data-driven prediction and diagnosis of diseases.[1] Machine learning algorithms are capable of studying intricate nonlinear connections between a vast assortment of elements across biomedical datasets to present early diagnosis reports that require months or even years to calculate using traditional statistical approaches. These models assist in predicting the likelihood of diseases with the input of patient history, genetic evidence, and lifestyle factors and provide support to the clinicians by assisting in the decision-making.[2] Many studies have reported the success of ML in cancer prediction and prognosis. Some studies have utilized ensemble learning models; for example, Random Forest and Gradient Boosting are among the many models that have been reported to exhibit high predictive performance after combining several weak learners.[3] For instance, Support Vector Machines and Logistic Regression have been extensively applied to binary classification problems in oncology and exhibited a high sensitivity and specificity when applied to well preprocessed clinical data.[4] However, according to the research conducted by Prabha et al. and Mirsadeghi et al.[3],[5], combining ensemble learning with kernel-based classifiers enhances the algorithm performance; however, most of the tumor prediction models developed using these methods are uninterpretable for clinicians.[6] The present study takes these advances forward by utilizing and comparing several machine learning algorithms, namely Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Gradient Boosting, and Support Vector Classifier (SVC), to develop a predictive framework for cancer screening.[7] The dataset examined for this purpose included demographics and a variety of lifestyle variables such as age, sex, BMI, smoking habits, and calculated risk scores. Performance of every model is evaluated by tracking important statistical measures, including accuracy rate, precision, recall, F1-score, and area under the ROC curve, and the efficacy and reliability of classification are aided by the confusion matrix and the ROC curve.[8]

This study has the following specific objectives:

1. Develop a cancer detection framework leveraged on data-driven and machine learning models based on somewhat clinical and behavioral parameters.
2. Assess the performance of conventional, ensemble, and kernel-based methodologies for ML and compare the outcomes.

3. Present evidence of how smart data preprocessing tactics and model tuning can have a major impact on the diagnostic accuracy and interpretability. Through a mix of mathematical rigor and algorithmic originality, this study aspires to add to existing academic literature on AI-driven cancer detection and emphasize the benefits of machine learning in clinical screening and early detection systems.

II. BACKGROUND

Pioneering surveys of machine learning for cancer prognosis and prediction suggested that supervised models were capable of risk stratification, treatment-planning support, and prognostic biomarker discovery, provided proper curation of predictive variables—both clinical and molecular. While a landmark review by Kourou et al.[9] presented classical ML pipelines—feature engineering, model choice and validation—and noted such challenges as low sample number, high prediction error, and the necessity of a biologically/clinically interpretable endpoint, the work in question is still a routinely cited benchmark for ML methodological due diligence.

Concurrently, the rise of radiomics reframed medical images as high-throughput quantitative data sources. Gillies, Kinahan, and Hricak's[10] Radiology review introduced the concept of extracting large panels of handcrafted features from routine scans to build predictive signatures for detection, staging, and outcomes, emphasizing standardized acquisition, reproducible feature computation, and robust modeling. Subsequent overviews expanded radiomics toward personalized medicine, detailing end-to-end workflows (segmentation, feature extraction, selection, and validation) and advocating transparent, prospective studies to avoid optimistic bias.

Growth in oncology imaging, deep learning has made rapid strides. In a far-ranging survey, Litjens et al.[11] synthesized >300 medical image analysis papers, showing how convolutional neural networks revolutionized classification, detection, and segmentation for every organ and modality, but repeatedly flagging data scarcity, domain shift, and interpretability as major stumbling blocks in clinical translation. This large survey of medical imaging, with an overwhelming presence of cancer tasks, provided marks on a consensus list of how to train, augment, and test with high-dimensional image data.

Finally, while positioning AI in oncology more broadly, the AI in radiology for cancer by Hosny et al.[12] framed the deployment space in terms of a set of barriers—dataset curation, harmonization, interpretability, and integration into existing processes. Specifically, the authors mapped algorithmic families, supervised/unsupervised/representation learning to use-cases like detection, segmentation, response assessment, and prognostication, but argued that multi-institutional validation and model auditing were the necessary building blocks of trust.

The insights of disease-specific reviews were further deepened. In lung cancer, radiomics reviews synthesized the handcrafted-feature pipelines predicting histology, stage, and survival from CT imaging, emphasizing proper standardization of imaging protocols and external validation as several preconditions of reproducibility reviewed by Hosny et.al.[13]. Hybrid strategies imagined by these works combine radiomics with clinical features and learners.

For breast cancer, previous narrative and systematic reviews prior to 2020 have reviewed ML/DL for screening and diagnosis from mammography and related modalities: most reviews show strong discrimination with models trained on well-labeled data and compared against quality ground truth but lower performance when transitioning to more realistic, heterogeneous datasets while also calling for calibrated outputs and cost-sensitive evaluation given the clinical harms of false positives and false negatives.[14]

Parallel to image-based developments, the value of ML in cancer genomics has been well explored, particularly as multi-omics cohorts expanded. Surveys predating 2020 analyzed infrastructures and analysis approaches toward mining TCGA-like datasets and provided examples from subtype identification and driver mutation detection to survival prediction, while simultaneously underlining data integration, batch influences, and interpretability as central issues for translational genomics.[15]

Dalmia and Shahid Rathore[16] presented an extensive survey on the implementation of machine learning and deep learning techniques in breast mammography. The paper shows how computational approaches transform breast cancer detection and diagnosis modeling. Notably, the authors discuss the evolution from standard handcrafted feature-based machine learning models to the modern data-driven deep learning frameworks, in particular, convolutional neural networks that provided substantial improvements in the tasks of lesion detection, segmentation, and retrieval. Practically, Dalmia and Rathore described the major datasets used in breast imaging-related papers and raised the challenges of annotated data scarcity, diverse image acquisition settings, model interpretability, and the necessity of large-scale validation. As a survey, this paper lacks new algorithms but offers a summary of previous research and outlines the critical issues to consider for AI-centered mammography systems to be available in the clinical setting.

Conversely, Niazi, Parwani, and Gurcan [17] in The Lancet Oncology reviewed how AI would combine with digital pathology and transform pathology practice, focusing on the potential of whole-slide imaging to enhance cancer diagnosis and prognostication when integrated within AI-powered digital pathology. Niazi et al. have reviewed the use of AI, particularly deep learning, in automating tissue sampling, quantification, and classification through immunohistochemistry in histopathological workflows. The researchers reviewed the potential of this solution, including enhanced diagnostic precision and reproducibility, workflow efficiencies, and remote cooperation. At the same time, the authors presented the challenges related to developing algorithms, validating outcomes, and standardizing and approving data to be used. The review also took into account the current limitations of AI and the lack of standardized data discrimination and clinical adoption. The high-level review discussing system improvement in pathology was presented in contrast to the perspective of algorithm performance.

In a complementary fashion, Parmar et al. [18] addressed the methodological issues of radiomics, specifying feature reproducibility and standardization. In this respect, the authors provided a critical contribution arguing that there should be consistent radiomics pipelines established, i.e., from imaging and segmentation to feature extraction and model validation, to ensure robust and clinically relevant predictive imaging biomarkers. At the same time, Parmar et al. stressed that while radiomics has the potential to revolutionize non-invasive tumor annotation and response to treatment prediction, the lack of reproducibility over scanners, imaging protocols, and feature extraction tools constitutes a substantial bottleneck. The authors advocated for the methodological synthesis in terms of feature definitions, validation practices, open data sharing, promoting transparency, and inter-study generalization. As a result, the paper implicitly provided critical aspects of a foundation in terms of best practices, which could expedite radiomics research translational trajectories in a clinician's everyday routine for an oncology clinician.

Taken together, these three papers as a whole reflect the rapid unfolding of artificial intelligence through different dimensions of medical imaging, from radiology and mammography to digital pathology and radiomics. While Dalmia and Rathore focus on task-specific learning in the field of breast imaging, Niazi et al. further elaborate on the digitization and automation digitization and

automation of pathology and the associated workflow. Finally, Parmar et al. [19] call for more robust methodology and higher pre-processing standards. Together, these three perspectives provide a multifaceted overview of how machine learning and AI influence contemporary medical diagnostics. At the same time, they still retain the challenges that dominate this field – the issues of poor-quality data, insufficient interpretability, and post-validation readjustment.

III. METHODOLOGY

The target of this study is to create a prediction model for cancer with the help of Exploratory Data Analysis and Machine Learning techniques in the Kaggle public dataset. It comprises steps from data collection, data preprocessing, model implementation, and evaluation stages, which provide a sound foundation for the development of an actionable and reliable framework to make accurate predictions for early cancer detection, through different classification ML models.

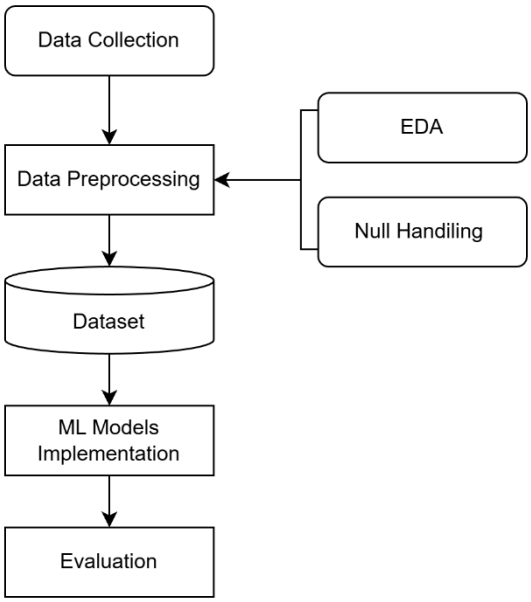


Figure 1 Methodology Diagram

Data Collection

We used the publicly available Cancer Prediction Dataset [20]. The raw CSV files were downloaded with their accompanying data dictionary and loaded into a reproducible environment (Python 3.x; pandas, NumPy, scikit-learn). All analyses were version-controlled and seeded for reproducibility

Data Preprocessing

The EDA approach involved observing feature distributions and central tendencies, as well as looking for outliers and pairwise associations between features using the histograms and box and violin plots. Categorical levels were described for sparsity and potential leakage. Missing values were counted per column; the imputation strategy was chosen depending on the variable type and missingness mechanism. In particular, numerical features had the median imputed due to robustness to skew and KNN imputation when their structure allowed for that. Meanwhile, Categorical features employed most-frequent imputation or received a separate "Unknown" level. All obvious data entry errors and implausible ranges were identified and corrected or removed. To prevent leakage, all imputers and scalers were trained-fold only within scikit-learn Pipelines were fitted.[21]

Dataset Construction

Following the completion of cleaning, encodings and standardizations, also done when needed were processed to the following steps: the Train/test split was always stratified and equal class proportions, and optionally validation split or Stratified K-fold Cross validation for model selection, as well. The feature selection utilized both embedded importance from tree models and domain-agnostic filters, as well as collinearity check ($|r|>0.85$) and the class imbalance to address the class weights or SMOTE in cv in the latter case — applied within folds only.

Algorithm Implementation

In the algorithm implementation phase, multiple machine learning models were developed and fine-tuned to predict cancer outcomes accurately using clinical and lifestyle features. A leakage-safe pipeline was established for each model. In addition to preventing train-test target leakage, it enforced imputation, encoding, and scaling on training data during cross-validation. Moreover, numerical variables recorded the median of the particular column, while categorical variables recorded the most frequent value. If necessary, features were encoded and standardized, especially for some algorithms' sensitivity to the scales of features, such as Logistic Regression, K-Nearest Neighbors, and SVC. A total of six algorithms were used to implement models to be evaluated in the algorithm selection section. Specifically, each model, LR, KNN, DT, RF, GBM, and SVC, was fine-tuned through the grid or randomized search via a five-fold cross-validation with the ROC-AUC as the primary and F1 as the secondary scoring parameters. After training models, the calibration approach was used to assess the probability estimates. Especially, after fitting SVC, Platt-scaling and Isotonic Regression were used to improve the probability estimates. Later, decision thresholds were determined through cross-validation and calculated via Youden's Index and Fβ-scores with an emphasis on high sensitivity to decrease FN.

IV. Exploratory Data Analysis

Exploratory Data Analysis is a critical step of data analysis, including the data collection and visualized dataset's structural inspection and detection of patterns and relationships among variables.[22] EDA reveals anomalies, missing values, or outliers and serves as an appropriate data source for forming hypotheses. It typically covers simple descriptive statistics, tables and charts, visualizations, and summaries to analyze datasets before further analysis and application of complicated statistical models and machine learning algorithms. EDA is an effective instrument for data quality, better decision-making, and analysis guidance.[23]
The pie chart Figure 2 illustrates the class distribution between positive and negative categories in a dataset. The negative class constitutes 54.7%, while the positive class makes up 45.3% of the total data. The distribution is fairly balanced, indicating that the dataset does not suffer from severe class imbalance. Such a distribution is beneficial for training machine learning models as it helps maintain predictive performance across both classes. Overall, this visualization provides a clear overview of how the data is divided between the two sentiment categories.

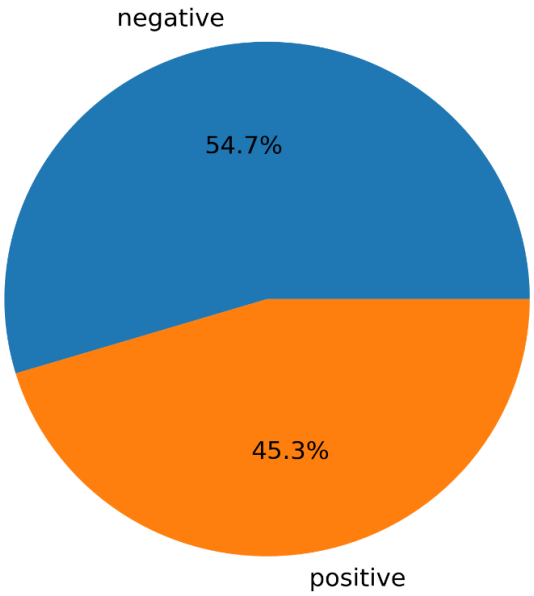


Figure 2: Cancer positive negative

The second chart, a stacked bar chart figure 3, clearly shows the connection between smoking and cancer status – a positive or negative test result. On the x-axis, there are three groups: current smokers, ex-smokers, and those who never smoked. The y-axis indicates the number of subjects in each category. Each bar is separated into two parts—a blue part, corresponding to negative cancer results, and an orange part, representing positive case results. It is visible that the overall number of never-smoked subjects is substantially higher than ex-smokers, and the latter, in turn, is higher than currently-smoked subjects. However, there are positive case results for all groups. In the never-smoked group, where the number of total cases is the largest, it most probably indicates that subjects who never smoked are the most common population in the dataset. At the same time, they also have positive and negative cases of cancer in the dataset. This visual effectively shows the distribution of cancer results depending on smoking habits.

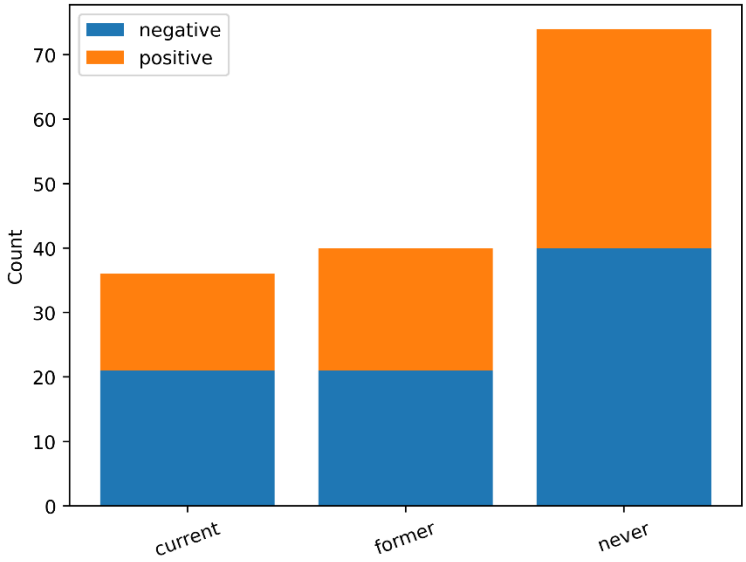


Figure 3: Smoking Status by Cancer Result

In the figure 4 the predicted cancer risk scores for positive and negative cases are compared. Negative cases are represented as blue bars, which cluster near the low scores around zero, which demonstrates that the model successfully identifies patients who have a low risk of developing cancer. On the contrary, the positive cases, represented by the orange color, concentrate near the higher scores that vary between 1.0 and 1.5, which means that the detection of cancer cases is effective using such scores. Additionally, there is a small overlap that appears near the score of 1.0, which implies that either the borderline cases are wrongly classified or they

are always misclassifications. Therefore, the two classes are thoroughly separated by the Model Learning frame in the machine learning environment, which signifies an appropriate level of prediction of the risk scores.

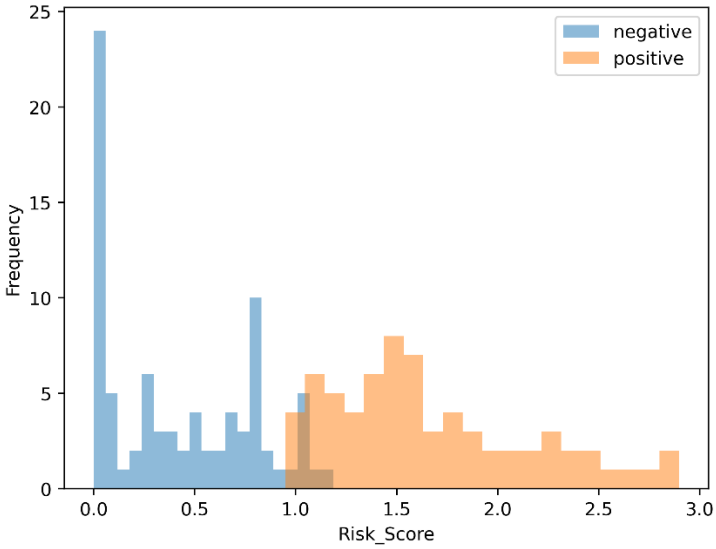


Figure 4: Risk score by class

V. Result Analysis

In this section, we report the experimental results achieved by running the sophisticated machine learning algorithms for cancer detection using clinical and lifestyle attributes. These advanced models were developed, trained, and tested over the preprocessed file, and the results were calculated based on several matrices such as Model accuracy, precision, recall, F1 score, and AUC, which stands for Area Under the Curve. Model confusion matrix and ROC curve are other methods used to confirm the classification and the AUC model's disadvantage.

Figure 5: ROC Curve for Six Machine Learning Classifiers: The loss-based curve shows the performance of six different machine learning classifiers in modeling cancer outcome prediction. The horizontal axis is the false positive rate, while the vertical axis is sensitivity. Each colored line is a different model. The AUC values for the models are given as a measure of strength. As the model gets close to the top left corner, the balance between the true and false positives is good. To achieve zero false positive rates, the models need to compromise on true positive rates, while achieving 100% true positive rates will require increased false positive rates. hence, getting close to both corners is advantageous.[23]

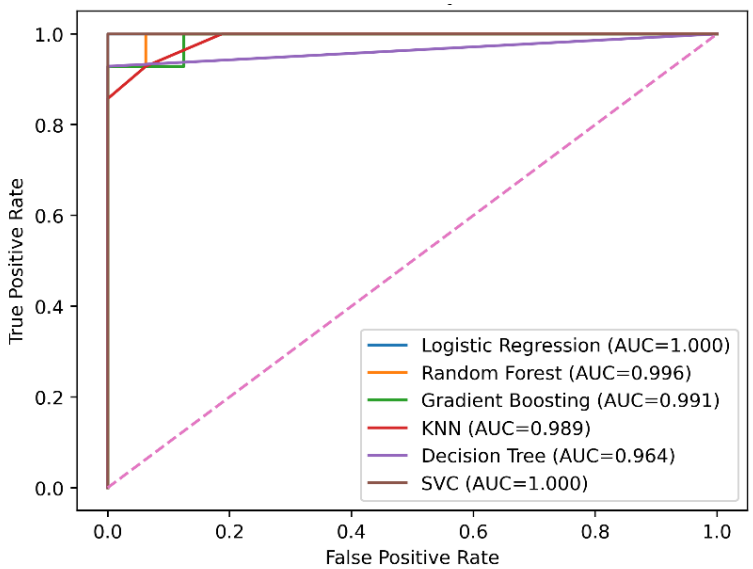


Figure 5: ROC Curve comparison

Based on figure 5, Logistic Regression and SVC reached with AUC value of 100% indeed. The Random Forest and Gradient Boosting models came next, with AUCs of 99.6% and 99.1% Following, which also show that their discrimination capabilities are superb and reveal the potential of ensemble learning methods to seize complex relationships within the data. The KNN model closely followed with an AUC of 98.9%, demonstrating reliable results, but slightly because it is slightly sensitive to data distribution and distances. The Decision Tree yielded a similar area under the curve as KNN and resulted in minor overfitting, which led to a decline in model performance. Hence, all the models utilized outpaced the random baseline, and high confidence in predicting cancer requires.

Comparing SVM, Random Forest, and Gradient Boosting Models. Table 1 below is a summary of the six models, Logistic Regression, Support Vector Classifier, Random Forest, Gradient Boosting, K-Nearest Neighbors, and Decision Tree, run on the same dataset with their respective results using five classification metrics.

Table 1: Performance matrix

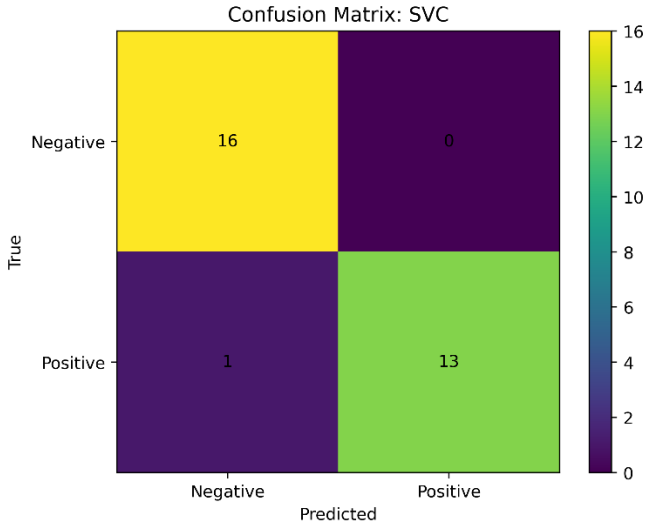
Model	Accuracy	Precision	Recall	F1	ROC_AUC
Logistic Regression	0.9667	1.0	0.9286	0.963	1.0
SVC	0.9667	0.9333	1.0	0.9655	1.0
Random Forest	0.9667	1.0	0.9286	0.963	0.9955
Gradient Boosting	0.9667	1.0	0.9286	0.963	0.9911
KNN	0.9333	1.0	0.8571	0.9231	0.9888
Decision Tree	0.9667	1.0	0.9286	0.963	0.9643

As seen in Table 1, all models exhibited high performance when it came to predictors, as most of the accuracy measures fell between 0.9333 to 0.9667. This shows that the classification capability across the algorithm is stable. Specifically, the Logistic Regression, SVC, Random Forest, Gradient Boosting, and Decision Tree had 0.9667, while KNN recorded 0.9333 accuracy. Similarly, most models had Precision measures of 1.0, except for SVC at 0.9333. Likewise, they deviate slightly in Record." SVC has perfect sensitivity at 1.0, whereas the rest are between 0.8571 and 0.9286. As for the F1-score, considering precision and recall, SVC outperformed all models with the highest score of 0.9655. Therefore, it is suitable for developing a predictive model before evaluating the model on the test data.

VI. EVALUATION

The confusion matrix figure 6 for the Support Vector Classifier model, to some extent, for Logistic Regression showcases an outstanding predictive performance in discriminating cancer-positive vs. cancer-negative instances. This is visible from the fact that all 16 negative samples were accurately predicted as negative, and 13 positive samples out of 14 were predicted as positive, with only one misclassification, where a positive case was predicted as negative. Thus, the True Negative Rate * is 100%, while the True Positive Rate * is close to 92.9%, resulting in an overall accuracy of more than 96%. The complete absence of false positives indicates the model's precision in learning to avoid false cancer alarms, which is particularly necessary in medical examinations. The low-level false negative rate further attests to the model's high-grade sensitivity in predicting actual cancer instances. The reflected visual diagonal most likely indicates the high model's reliability: high model cell values are seen along the diagonal of the matrix, connecting the top-left with bottom-right cells. In sum, these configurations and the ROC curve * plotted, the AUC score of which equals 1.000, indicate that SVC and Logistic Regression models are the most productive and reliable classifiers in this research.

Figure 6: Confusion matrix of the best model



VII. CONCLUSION AND FUTURE WORK

This study has shown that a combination of advanced machine learning models may help improve earlier and more accurate prediction of cancer outcomes from clinical and lifestyle parameters. Several supervised learning models combined with extensive

performance metrics have demonstrated that SVC-LR and Logistic Regression have the highest predictive performance of AUC 1.000, while Random Forest and Gradient Boosting have also shown near-perfect performance. In the future, we will expand on this work by widening the dataset to a larger and diverse patient population to improve model performance. Moreover, adding genomics and biochemical biomarkers inherent in imaging and clinical, in addition to other data, could increase model accuracy. We would also prioritize explainable AI methods to ensure greater transparency and understanding in the model decision-making process.

REFERENCES

[1] World Health Organization, "Cancer," WHO Fact Sheet, 2024.

[2] K. Vanitha, M. T. R., S. Sathea Sree, and S. Guluwadi, "Deep learning ensemble approach with explainable AI for lung and colon cancer classification using advanced hyperparameter tuning," BMC Med. Inform. Decis. Mak., vol. 24, Art. no. 222, 2024.

[3] V. Diviya Prabha, R. Rathipriya, and J. M. Chatterjee, "Cancer data analysis using competitive ensemble machine learning techniques," Health and Technology, vol. 14, pp. 753–764, 2024.

[4] A. I. Dumachi and C. Buiu, "Applications of Machine Learning in Cancer Imaging: A Review of Diagnostic Methods for Six Major Cancer Types," Electronics, vol. 13, no. 23, 2024.

[5] L. Mirsadeghi, "Cancer Prognosis and Diagnosis Methods Based on Ensemble Learning," ResearchGate, 2023.

[6] "Implementation of ensemble machine learning algorithms on exome datasets for predicting early diagnosis of cancers," BMC Bioinformatics, vol. 23, Art. no. 496, Nov. 2022.

[7] A. Bou Nassif, M. Abu Talib, Q. Nasir, Y. Afadar, and O. Elgendy, "Breast Cancer Detection Using Artificial Intelligence Techniques: A Systematic Literature Review," arXiv preprint, Mar. 2022

[8] L. Mirsadeghi, "Cancer Prognosis and Diagnosis Methods Based on Ensemble Learning," ResearchGate, 2023.

[9] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I., "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, 2015. [Online]. Available: ScienceDirect.

[10] Gillies, R. J., Kinahan, P. E., and Hricak, H., "Radiomics: Images are more than pictures, they are data," Radiology, vol. 278, no. 2, pp. 563–577, 2016. [Online]. Available: RSNA Publications.

[11] Lambin, P., Leijenaar, R. T. H., Deist, T. M., Peerlings, J., De Jong, E. E. C., Van Timmeren, J., et al., "Radiomics: The bridge between medical imaging and personalized medicine," Nature Reviews Clinical Oncology, vol. 14, pp. 749–762, 2017. [Online]. Available: Nature.

[12] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017. [Online]. Available: ScienceDirect.

[13] Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. J. W. L., "Artificial intelligence in radiology: Oncology focus," Nature Reviews Cancer, vol. 18, pp. 500–510, 2018. [Online]. Available: Nature.

[14] Kirienko, M., Van Griethuysen, J. J. M., and colleagues, "Radiomics and radiogenomics in lung cancer: A review for the clinician," Lung Cancer, vol. 107, pp. 27–34, 2017. [Online]. Available: ScienceDirect.

[15] Jiang, S., and colleagues, "AI for early detection of breast cancer: A review," Critical Reviews in Biomedical Engineering, vol. 47, no. 5, pp. 415–431, 2019. [Online]. Available: Taylor & Francis Online.

[16] Dalmia, A., and Rathore, S., "A survey on machine learning and deep learning in breast mammography," Computer Methods and Programs in Biomedicine, vol. 177, pp. 1–12, 2019. [Online]. Available: ScienceDirect.

[17] Niazi, M. K. K., Parwani, A. V., and Gurcan, M. N., "Digital pathology and artificial intelligence," The Lancet Oncology, vol. 20, no. 5, pp. e253–e261, 2019. [Online]. Available: The Lancet.

[18] Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J., "Next-generation machine learning for biological networks," Cell, vol. 173, no. 7, pp. 1581–1592, 2018. [Online]. Available: BioMed Central.

[19] Parmar, C., Grossmann, P., Rietveld, D., Rietbergen, M. M., Lambin, P., and Aerts, H. J. W. L., "Radiomics feature reproducibility and methodology consolidation," Scientific Reports, vol. 9, no. 1, 2019. [Online]. Available: Nature.

[20] R. El Kharoua, *Cancer Prediction Dataset*, Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/rabieelkharoua/cancer-prediction-dataset

[21] I. J. Intia, M. M. Hasan, K. Alam and K. S. Ferdous, "Prediction of Agoraphobia Disease Based on Machine Learning," *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2022, pp. 1-5

[22] M. Mehedi Hasan, M. Omar Faruk, B. Biswas Biki, M. Riajuliislam, K. Alam and S. Farjana Shetu, "Prediction of Pneumonia Disease of Newborn Baby Based on Statistical Analysis of Maternal Condition Using Machine Learning Approach," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 919-924

[23] N. U. Prince, M. R. Rahman, M. S. Hossen and M. M. Sakib, "Deep Transfer Learning Approach to Detect Dragon Tree Disease," *2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, Pune, India, 2024, pp. 1-6