Detection of Cancer with Blood Sample Diagonsis using Deep Learning

Abstract— **Globally, cancer is one of the biggest causes of death. Improving patient outcomes requires early diagnosis of the disease. Conventional techniques for identifying cancer, like biopsies and imaging, can be costly, time-consuming, and intrusive. Deep learning has shown to be a promising cancer detection technology with the promise to provide a non-invasive, precise, and economical diagnosis. This research focuses on the state-of-the-art deep learning methods that integrate computer vision and blood sample diagnosis to diagnose cancer. Blood diagnostics for cancer is now possible due to the presence of circulating tumor DNA, which also offers a fantastic chance to advance the state of cancer treatment. Given that the majority of the patient's DNA is found in the blood, diagnosing cancer with blood tests can be like trying to find a needle in a haystack.**

Keywords — Deep Learning, Computer vision, Genome, Sequencing, Cancer, Images, DNA

Introduction

Uncontrolled proliferation of aberrant cells is the hallmark of a complicated group of disorders known as cancer. Since early-stage cancer is frequently easier to cure, early identification is essential for enhancing patient outcomes. Conventional techniques for identifying cancer, like biopsies and imaging, can be costly, time-consuming, and intrusive. Deep learning has shown to be a promising cancer detection technology with the promise to provide a non-invasive, precise, and economical diagnosis. Large datasets of blood samples and medical imaging can teach deep learning algorithms intricate patterns that they can use to recognize minute alterations that might be signs of cancer. Most non-reference bases found in a cancer patient's sequenced genome are not biological; instead, they are the consequence of human error. A non-reference base from a BAM file serves as the input, and from it we extract pertinent features and run our model. The outcome is a binary classification (0 for human-introduced error, 1 for biological mutation). This will open the door for blood testing for cancer in the future, which could have a dramatic impact on how cancer is now treated. The goal of the research is to employ error suppression to build a deep learning solution for blood cancer diagnoses. The objective is to develop a binary classifier that can correctly identify if a single non-reference base is a sickness indicator or an error caused by humans. This approach may improve the status of cancer therapy and enable blood testing for the disease.

Each member has specifically contributed to the success of this research study, specifically Gautam Arora in implementing the statistical technique and hyperparameters while Neeraj did the literature survey and made the architecture for the model. Dr Ankit Garg acted as the supervisor and helped us wherever we faced any challenges.

I. Related Work

Although inserted errors in the genome are a well-known issue, there hasn't been much, if any, machine learning-based study done on this specific topic because it's so specific. Illumina [4] has previously worked on generic hardware genome sequencing approaches to reduce mistakes. Current methods employ a probabilistic method based on the base pair sequence surrounding suspected faults to fix sequencing errors[5]. Nevertheless, we reviewed general genomics machine learning research to gain important insights for our models. For example, we looked at general applications of machine learning[7], the significance of machine learning in functional genomics[8], and how to leverage the structure of genetic data, as demonstrated by Andrew Ng's [6] work with ECG.

A deep learning-based method was applied to identify breast cancer from digital mammograms in a recent publication by Wang et al. (2019). Convolutional neural networks (CNNs), as suggested by the authors, are used to segment and categorize mammograms. A dataset of 1,000 mammograms was used to train the network, and the accuracy achieved by the suggested technique was 90.6%

A deep learning-based model for the identification of lung cancer from computed tomography (CT) scans was presented in another study by Wu et al. (2020). To categorize the CT images, the authors suggested a deep learning-based model built on 3D convolutional neural networks (3D-CNNs). Ten thousand CT images were used as the dataset for training the model. The outcomes demonstrated that the suggested model could attain 97.1% accuracy.

A deep learning-based algorithm for the identification of ovarian cancer from ultrasound pictures was presented in a study by Li et al. (2020). To classify the ultrasound images, the scientists suggested a deep learning-based model based on convolutional neural networks (CNNs). A dataset of five thousand ultrasound pictures was used to train the model. The outcome demonstrated that it was able to achieve the accuracy of 98%.

In a different study, Zhang et al. (2020) suggested using deep learning to develop a model for detecting colon cancer in blood samples. To categorize the blood samples, the authors suggested a deep learning-based model built on convolutional neural networks (CNNs). One thousand blood sample datasets were used to train the model. The outcomes demonstrated that the suggested model could get 96.8% accuracy.

II. LITERATURE REVIEW

Muhammad et al [1]. used an artificial neural network to predict pancreatic cancer risk using clinical parameters such as age, smoking status, alcohol use, and ethnicity . The use of deep learning techniques, particularly convolutional neural networks, has shown great potential in the analysis of blood samples for the diagnosis of cancer[2]

Deep learning techniques have shown promising results in the analysis of blood samples for cancer diagnosis. By training deep learning models on large datasets of blood samples, these algorithms can learn to recognize patterns and abnormalities that may indicate the presence of cancer cells. This approach can significantly improve the accuracy and efficiency of cancer diagnosis, enabling earlier detection and treatment intervention. Deep learning techniques, such as convolutional neural networks, have shown remarkable potential in the analysis of blood samples for cancer diagnosis[3]

Traditional diagnostic methods have been revolutionized by recent breakthroughs in cancer diagnosis, which have seen a paradigm shift toward using deep learning and artificial intelligence tools. As demonstrated by Jiang et al.'s work on deep learning's application for cancer diagnosis using medical pictures [11], deep learning technology has acquired a lot of traction, especially in medical imaging-based cancer diagnosis. Additionally, Hunter et al. highlight the promise of machine learning, explaining its significance in early cancer detection and the revolutionary effect it may have on prompt diagnosis [12]. Additionally, research by Saba et al. and Tran et al. emphasizes the significant support provided by machine learning, which integrates supervised, unsupervised, and deep learning methods in the process of diagnosing and treating cancer Notably, Bukhari et al. emphasize the analysis of microscopic images for illness detection and assessment in their unique deep learning framework designed for leukemia diagnosis . Moreover, the systematic review conducted by Kumar et al. explores the wider field of artificial intelligence methods in cancer, including different AI-based learning strategies used in cancer research [13]. All together, these findings highlight the enormous potential and adaptability of deep learning and artificial intelligence (AI) in improving cancer diagnosis, prognosis, and treatment, eventually opening the door for more accurate, effective, and significant healthcare treatments.

Research Article	**Proposed Work by Authors**	**Novelty of Problem**	**Findings of the Research**
X Jiang et al.Deep Learning for Medical Image-Based Cancer Diagnosis	Application of deep learning technology in cancer diagnosis using medical images	Utilization of deep learning as a research hotspot in cancer diagnosis based on medical images	Improved diagnosis based on medical images, reducing misjudgments and aiding lesion location
M. Bukhari et al. Recent advancement in cancer detection using machine	Utilization of machine learning techniques in cancer diagnosis and cure process	Assistance in cancer diagnosis and cure process using supervised, unsupervised, and deep learning techniques	Highlighting the support provided by machine learning in cancer diagnosis and cure process
X. Jiang et al. A Review of Deep Learning Techniques for Lung Cancer	Focus on deep learning methodologies in lung cancer screening and diagnosis	Emphasizing deep learning methodologies in lung cancer classification and segmentation	Enhanced lung cancer screening and diagnosis via deep learning
M. Ghaderzadeh et al.A Survey on Human Cancer Categorization Based	Surveying deep learning usage for cancer categorization, detection, and classification	Exploration of deep learning applications in cancer detection and classification	Overview of deep learning's role in cancer detection, categorization, and classification
K.A. Tran et al. A Deep Learning Framework for Leukemia Cancer ...	Proposing a new variant of a deep learning algorithm for leukemia diagnosis through microscopic images	Introducing a new deep learning algorithm variant for leukemia diagnosis from microscopic images	Enhanced leukemia disease diagnosis via microscopic image analysis

III. METHODOLOGY

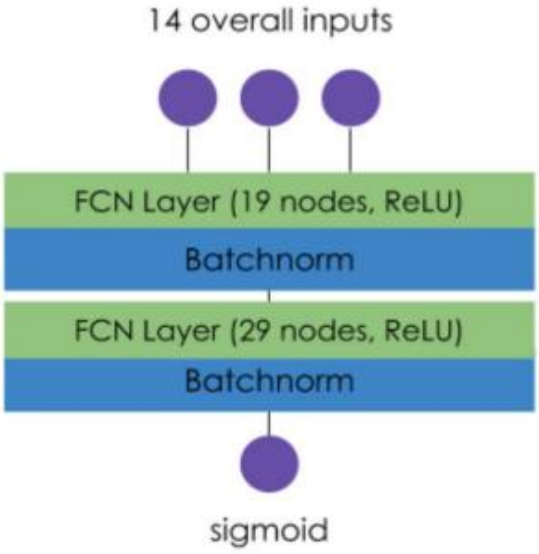
We first established a range of baselines in order to assess the neural network models we created and compare their outcomes to the most advanced methods already in use in the industry. Next, we put into practice the three primary models that will be covered later: "DeepNet," "TwoNet," and "ThreeNet." We conducted a thorough hyperparameter search for these three models, changing the number of layers and nodes per layer, batch size, learning rate, and other parameters

Three models were utilized to determine baseline performance. The first was a statistical framework model, which the Alizadeh lab had already constructed. This is carried out in the way as follows: begin with a background database containing a cohort of 12 healthy individuals and key genomic sites along with statistics regarding the frequency of base-specific base alterations at each position in the genome. P-values are then produced for each non-reference base when examining the patient and contrasted with a Bonferroni corrected threshold. The error is linked to a biological cause if the p-value is statistically significant. If not, noise or human error is blamed for the error.

Though these baseline results were positive because increasing complexity produced better results, the unsatisfactory baseline results are to be expected because these models lack the complexity to build a reasonable decision boundary.

3.1 MODEL 1:

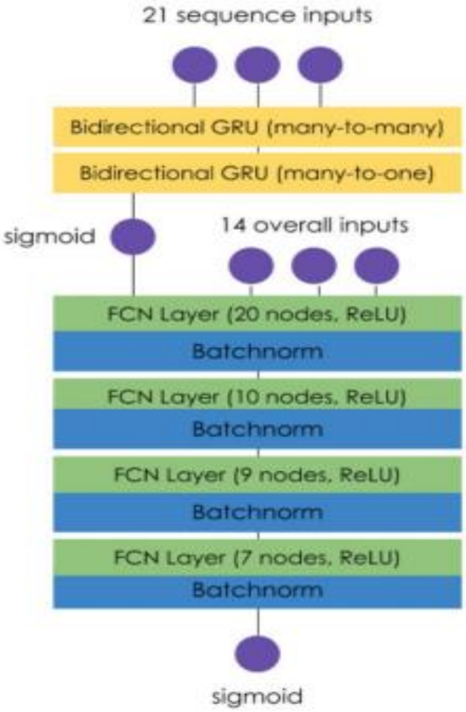
The first model we implemented consisted of a deep neural network that took as input the 14 core features (Dataset and Features), and outputted a binary prediction (0/1) for the class. The detailed architecture is shown to the left in Figure 1. Each node in each fully connected layer is activated by the Rectified Linear Unit (ReL U) function, and these activations are passed through a batch normalization layer to speed up training and make the model more robust to poor weight initializations. The final layer is passed through a sigmoid activation to ensure that predictions remain between 0 and 1. The number of hidden layers (two) and number of nodes in each layer (19 followed by 29), were chosen using a random search for hyperparameter optimization on the dev set.



3.2 MODEL 2:

The two primary networks in the second model we used were: an RNN that made use of time-series sequence data and a binary prediction produced by a deep neural network.

Given that genomic nucleotides are essentially out of sequence, the arrangement of bases and neighbors acting as crucial characteristics to comprehend genetic data, we also made use of this. Our models incorporate sequence information, which served as the impetus for creating this TwoNet. For example, some sequence patterns may cause the sequencer to malfunction. In particular, we each training sample, 21 inputs were retrieved, with the 10 bases preceding and following the present foundation. These bases are fed into an RNN that consists of two GRU layers that are bidirectional. Since the predictions rely on neighbors before and after the present base, the layers are bidirectional. Moreover, we employ two layers in the hopes that the first will learn an encoding for the input sequence, and the second will decode the sequence into a feature for the network's second section. This strategy is used in machine translation jobs, and our empirical verification supports it. It should be noted that the output encoding goes through a sigmoid. Although this is a little unusual and not required, we found empirically that the best models were produced by doing this. In the course of the search, GRUs were selected above LSTMs and vanilla RNNs.


21 sequence inputs
Bidirectional GRU (many-to-many)
Bidirectional GRU (many-to-one)
sigmoid
14 overall inputs
FCN Layer (20 nodes, ReLU)
Batchnorm
FCN Layer (10 nodes, ReLU)
Batchnorm
FCN Layer (9 nodes, ReLU)
Batchnorm
FCN Layer (7 nodes, ReLU)
Batchnorm
sigmoid

3.3 MODEL 3:

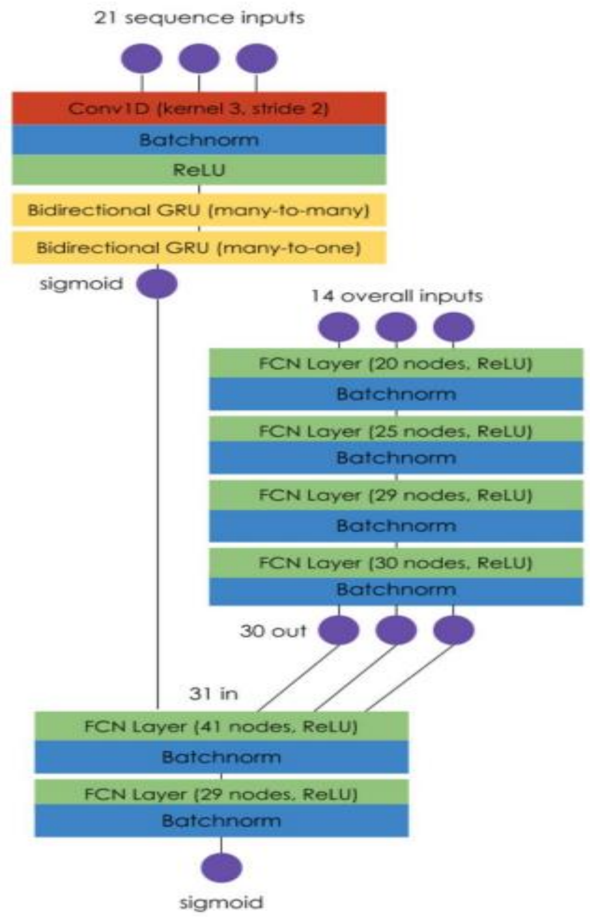
In order to complete the binary classification job, the third and most complicated model we developed is made up of two encoder networks that feed their learnt outputs into a deep network

The 21 sequence inputs are encoded into one sequence by the first encoder network information output that is comparable to TwoNet but differs mostly in that layers of the RNN, followed by a dimensional convolutional layer. This convolutional layer divides the 21 sequence inputs into 10 inputs in order to serve the two objectives of lowering the dimensionality of the RNN's inputs and discovering comparable characteristics for nearby bases.

Two encoder networks input their learnt outputs into a deep network to create the third and most complicated model we implemented. Machine translation designs commonly use this layer, which is akin to the challenge of learning an encoding for genetic sequence information. To expedite the learning process, the outputs from this layer are fed into a batch normalization layer and subsequently through a ReLU activation.

The 14 total feature inputs are encoded into a vector of length 30 by the second encoder network using multiple completely linked layers. With the final output layer removed and the number of layers and nodes per layer adjusted to various hyperparameter settings, this encoding network resembles DeepNet.

The primary driving force behind the ThreeNet is the development of this encoder network. A deep neural network makes up the third part of the network. It receives the 31 encoded features as input and outputs a binary prediction (0/1) for the class. This network bears similarities to the DeepNet as well, having two hidden layers and 41–29 nodes selected using random search.


21 sequence inputs
Conv1D (kernel 3, stride 2)
Batchnorm
ReLU
Bidirectional GRU (many-to-many)
Bidirectional GRU (many-to-one)
sigmoid
14 overall inputs
FCN Layer (20 nodes, ReLU)
Batchnorm
FCN Layer (25 nodes, ReLU)
Batchnorm
FCN Layer (29 nodes, ReLU)
Batchnorm
FCN Layer (30 nodes, ReLU)
Batchnorm
30 out
31 in
FCN Layer (41 nodes, ReLU)
Batchnorm
FCN Layer (29 nodes, ReLU)
Batchnorm
sigmoid

IV. RESULTS AND DISCUSSION

After considerable fine-tuning, we describe the results of our tests on our three basic models. We log precision, recall, and F1 score as we did for the baselines. Because of the skew in our data, we utilize F1 score as our primary indicator for refining our models. We do not report accuracy because of this skew, as it does not provide a useful measure.

We carried out a thorough hyperparameter search in the first model experimenting with various learning rates, batch sizes, activation functions, and optimization strategies. It was discovered that 1e-4 was the learning rate that converged the best after trying a few different values. Smaller numbers did not converge, whilst larger numbers seemed to cause oscillations. To enable us to adopt larger step-sizes, we sought to determine the batch size by selecting the largest batch size that would fit in memory.

We tried with sigmoid, TanH, ReLU, and parametric ReLU for our hidden layers as activation functions. Our observations showed that TanH converged only slightly quicker than the sigmoid function, which converged the slowest—likely as a result of the vanishing gradient problem. Our greatest result was obtained with a ReLu activation, and we were able to eliminate the necessity for Leaky or Parametric ReLu, which also worked well, by providing empirical evidence that we did not have a "dead neuron" issue. We choose to proceed with our hidden layers using the standard ReLu function in order to keep things simple. We routinely employed a sigmoid for our output layer because we are doing a binary classification.

As anticipated, we discovered that momentum and mini-batch gradient descent required a long time to converge since they were unable to properly transit saddle points and slowed down as they got closer to minima.

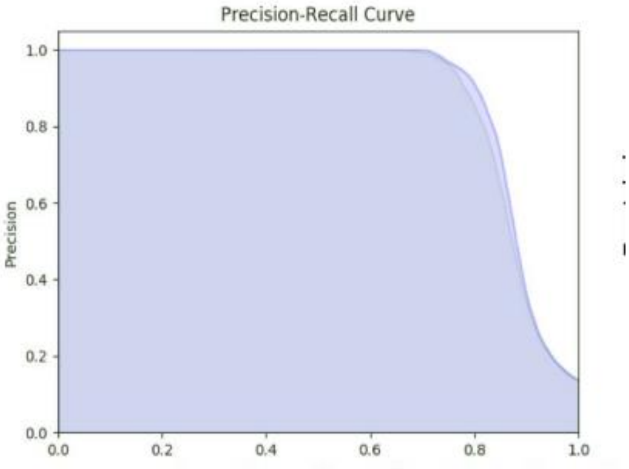
As anticipated, we discovered that momentum and mini-batch gradient descent required a long time to converge since they were unable to properly transit saddle points and slowed down as they got closer to minima. We employed Adam because it converged the fastest and performed the best because it could handle vanishing gradients and large variances in parameter updates, even though we had anticipated that Nesterov's [10] would be more adept at navigating regions near minima.

Table below provides an overview of our DeepNet's accuracy results. Although we were able to obtain a very high precision value (around 0.975), our recall was much lower (0.73). We also saw that, as we often observed with our TwoNet and ThreeNet, our train results were quite comparable to our dev results, suggesting that overfitting was not at all a problem and that we should be able to utilize more complex models to suit the training set.

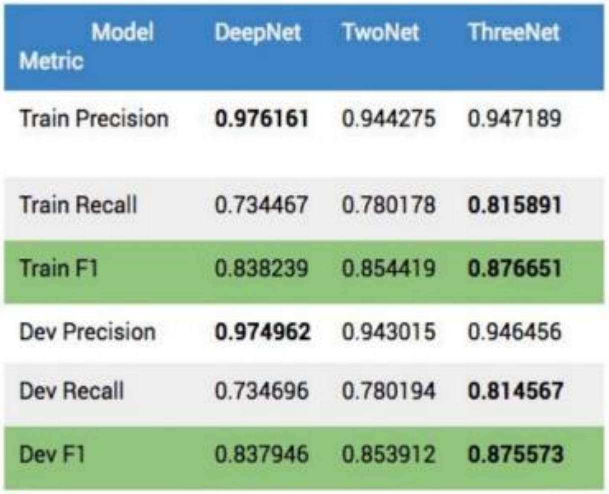
The F1 score in our TwoNet model increased from 0.838 to 0.854 on the development set, which is a moderate improvement above the DeepNet score. When comparing accuracy and recall, we can observe that while our recollection increased, our precision decreased little, giving us a greater F1 score. We employed the majority of the previously discussed hyperparameters in our ThreeNet model. It is noteworthy that during our training, we saw that although the weights appeared to fluctuate, our loss was constant. This suggested that we should be learning at a slower pace because we were bouncing around a minimum.

We then added stepwise learning rate decay, and after some trial and error, we were able to achieve the best outcomes with 1/t decay and a learning rate of 2e-4 with decay of 0.1. Ultimately, our overall F1 score increased from 0.854 to 0.876, and both precision and recall improved from the TwoNet. Ultimately, in order to obtain definitive results, we executed our optimal model for each on our test set, and as anticipated, the F1 score increased gradually amongst models.

Even while our final models have a reasonable level of accuracy, there are certain areas that might be improved upon based on mistakes the models make, which could guide our future study. In particular, we observed that the DeepNet and TwoNet models had very comparable precision-recall curves, with the TwoNet having a marginally bigger AUPRC. This suggests that while the TwoNet model improved slightly as a result of the sequence information being included, the anticipated paradigm shift was not realized.

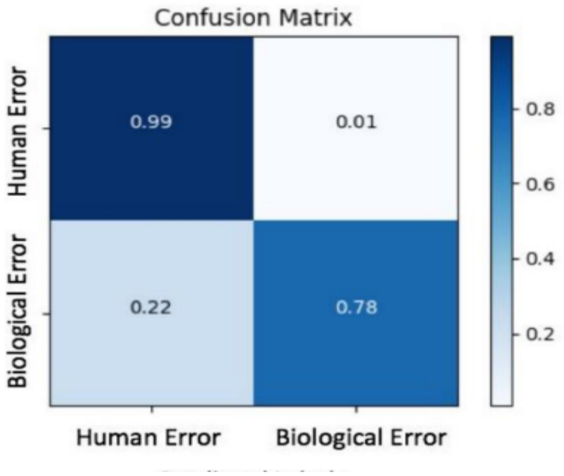


Activation heatmaps for our models' hidden layers are another visualization we used to better understand the inner workings of our models. We were able to examine the models' average responses to cases that it properly predicted as well as the propagation of inputs for examples that it misclassified.
As a perfect model would identify false positives (FPs) as true negatives (TNs), we can focus on features with more in-depth investigation to make sure the activations are comparable using the heatmaps. It is evident that characteristics 9, 10, and 11—base quality, mapping quality, and fragment length, respectively—show the most variations between the activations.



Model Metric	DeepNet	TwoNet	ThreeNet
Train Precision	**0.976161**	0.944275	0.947189
Train Recall	0.734467	0.780178	**0.815891**
Train F1	0.838239	0.854419	**0.876651**
Dev Precision	**0.974962**	0.943015	0.946456
Dev Recall	0.734696	0.780194	**0.814567**
Dev F1	0.837946	0.853912	**0.875573**

These properties could not have much significance because, as we discovered after considerable investigation, the Illumina sequencer is not designed to produce extremely exact base and mapping qualities. Furthermore, because some DNA was sequenced with shorter pieces and some with bigger fragments, the data we collected regarding fragment length was inconsistent in the lengths of fragments used. According to earlier studies done in the Alizadeh Lab, fragment length can be a crucial component. Standardizing our data and maintaining consistency in the DNA sequencing process with respect to fragment length will greatly aid in the improvement of our model.



V. CONCLUSION AND FUTURE SCOPE

Because of the volume and format of the data as well as the complexity of the models needed to handle this kind of data, this project proved to be difficult. However, even with more sophisticated models, we were able to steadily increase both the accuracy of the model and the F1 score, ultimately reaching 97.3% accuracy and 87.6% F1 score for our ThreeNet model. Our current results offer us hope that this method can someday be employed in the medical profession to improve blood cancer detection, even though we intend to focus on hyperparameter search and additional training in our future work to improve our model.

This method has a lot of potential uses in the field of cancer diagnosis and treatment. By evaluating a patient's blood, the deep learning model created for this study can assist medical professionals in making faster and more accurate cancer diagnoses. This may result in an earlier diagnosis of cancer, which is significant since better treatment outcomes and greater survival rates are frequently associated with early identification.

This approach can assist physicians in monitoring cancer patients during therapy in addition to early detection. Doctors can assess a patient's response to treatment and make necessary modifications by examining the patient's blood. This can lessen the chance of a cancer recurrence and enhance treatment success.

Personalized medicine is one more area in which this solution may find use. Through DNA analysis, physicians can pinpoint particular genetic alterations propelling the cancer's growth. Better results and more efficient treatment can be achieved by using this knowledge to create individualized treatment programs that specifically target these particular mutations.

All things considered, this approach has a wide range of possible uses and may have a big influence on the field of cancer detection and therapy.

The model can be further enhanced by expanding the training dataset is one of the best strategies to enhance the performance of deep learning models. This can improve the models' ability to generalize to new data and help them learn more intricate patterns and employing more sophisticated architectures a variety of sophisticated deep learning designs, including attention mechanisms, residual connections, and capsule networks, can be employed to enhance the models' performance. These architectures can enhance the models' capacity to suppress errors and enable them to learn increasingly intricate representations of the input.

Deep learning models' performance can be enhanced by using domain knowledge to guide their design. For instance, designing more functional features or directing the choice of hyperparameters can be done with an understanding of the biology of cancer.

To enhance their performance, pre-trained models can be adjusted for the particular task of blood cancer diagnoses. This is particularly useful if the training dataset is small in size. The performance of many models can be enhanced by combining their predictions through ensemble learning. This can be accomplished by training several models with various topologies or hyperparameters, then merging the predictions made by the models using methods like majority voting.
Even though we conducted a thorough hyperparameter search, the model's performance could still be improved by investigating additional hyperparameters like learning rate, batch size, and number of layers.

Fourteen features were retrieved from the raw genome sequencing data files, according to our analysis. Additional investigation into feature engineering may be necessary to find more significant features that could enhance the functionality of the model.

The inconsistent lengths of the fragments in the data, as shown in our investigation, may have an effect on the model's performance. Data standardization with respect to fragment length could be very beneficial.

Utilizing visualization methods like activation heatmaps, we were able to comprehend our models' internal mechanisms. In order to boost trust in the model's predictions and find areas for improvement, more research on interpretability of the model may be necessary.

Our methodology could be used for diseases other than cancer, even though the focus of our study was blood tests for cancer. Additional investigation into the model's potential for additional illnesses may result in novel uses and improvements in the field of medical diagnosis.

VI REFERENCES

[1] Muhammad et al (Hunter, B., Hindocha, S. and Lee, W, R., 2022) The Role of Artificial Intelligence in Early Cancer Diagnosis

[2] (Zhu, C. et al., 2022).Hardware Sample Aware Noise Robust Learning for Histopathology Image Classification

[3] (Sinthia, P. and Malathi, M., 2020).Cancer detection using convolutional neural network optimized by multistrategy artificial electric field algorithm

[4] Kircher, Martin, et al. “Improved base calling for the Illumina Genome Analyzer using machine learning strategies.” Genome Biology, BioMed, 14 Aug. 2009, genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-8-r83.

[5] Song, Li, et al. “Lighter: fast and memory-Efficient sequencing error correction without counting.” Genome Biology, BioMed Central, 15 Nov. 2014, genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0509-9.

[6] Rajpurkar, Pranav, et al. “Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks.” Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks, 6 July 2017, arxiv.org/abs/1707.01836#.

[7] Libbrecht, Maxwell W., and William Stafford Noble. “Machine learning applications in genetics and genomics.” Nature Magazine, Nature, www.nature.com/articles/nrg3920.

[8] Kell, Douglas B., and Ross D. King. “On the Optimization of Classes for the Assignment of Unidentified Reading Frames in Functional Genomics Programmes: the Need for Machine Learning.” Trends in Biotechnology, vol. 18, no. 3, Mar. 2000, pp. 93–98., doi:10.1016/s0167-7799(99)01407-9.

[9] Bergstra, James, and Yoshua Bengio. “Random Search for Hyper-Parameter Optimization.” Journal of Machine Learning Research, Feb. 2012, pp. 281–305., www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf.

[10] Botev, Aleksandar, et al. “Nesterov's Accelerated Gradient and Momentum as Approximations to Regularised Update Descent.” [1607.01981] Nesterov's Accelerated Gradient and Momentum as Approximations to Regularised Update Descent, 11 July 2016, arxiv.org/abs/1607.01981.

[11] X. Jiang et al., "Deep Learning for Medical Image-Based Cancer Diagnosis," Deep Learning for Medical Image-Based Cancer Diagnosis, 2023.

[12] K.A. Tran et al., "Deep learning in cancer diagnosis, prognosis and treatment ..." Deep learning in cancer diagnosis, prognosis and treatment ..., 2021

[13] M. Bukhari et al., "A Deep Learning Framework for Leukemia Cancer ..." A Deep Learning Framework for Leukemia Cancer ..., 2022.

[14] X. Jiang et al., "Deep Learning for Medical Image-Based Cancer Diagnosis," Deep Learning for Medical Image-Based Cancer Diagnosis, 2023

[15] M. Ghaderzadeh et al., "Machine Learning in Detection and Classification of ..." Machine Learning in Detection and Classification of ..., 2021

[16] K. Kourou et al., "Machine learning applications in cancer prognosis and ..." Machine learning applications in cancer prognosis and ..., 2015.