Abstract

The diagnosis of blood cancers presents a significant challenge due to the diseases' complex nature and the often-subtle changes in blood cell morphology. Conventional techniques of diagnosis involve reliance on pathologists for a manual evaluation of blood samples using a lens. Nonetheless, the process of using this approach is tiresome and there is a possibility of human interference. The introduction of machine learning, especially deep learning, to medical diagnosis has changed the way blood cells are identified and distinguished, as well as blood cancers noted, from other types of cells.

This study examines the possibility of using Random Forest, SVM, Gradient Boosting, as well as deep learning CNN, models for the classification of eight types of blood cells. The models were adjusted for effectiveness and while all of them performed well, the CNN model presented a high possibility for image-based medical diagnosis and to identify structural multiformity of blood cells. Techniques used includes data augmentation, including rotation and zoom to minimize overfitting and the subsequent results reflected on accuracy, precision, recall, AUC. Moreover, the research also employed the use of ensemble learning models, employing stacking the hyperparameter tuned-SVM and Random Forests. Logistic regression used in this study as a meta-classifier in the proposed ensemble model performed better in terms of accuracy and stability of both the test and validation datasets. These findings of this research suggest that enhanced machine learning algorithms can enhance the potential of blood cancer diagnosis which may in turn reduce mortality rates and enhance patient’s survival chances. At last, the given work aims to showcase those transforming diagnostics with machine learning is the key to the future of detecting blood cancer specifically. Through enhanced machine learning techniques and data augmentation, the study lays the foundation on how to create effective diagnostic tools for blood cancer that could bring positive results in the treatment process.

1. Introduction

In the complex world of medical diagnoses, blood cancers present a formidable challenge. Most of these diseases hamper the normal production of blood cells within the bone marrow and often progress without being detected until they reach an advanced stage. Of these, the aggressive types have an overgrowth of abnormal white blood cells crowding out the healthy cells, leading to serious health complications. The critical task of diagnosing blood cancers—where early detection can significantly impact patient outcomes—requires timely and accurate diagnosis. Timely and accurate diagnoses are paramount.

Decades ago, most of the diagnosis of blood cancer depended on the fastidious work of pathologists working for hours at the microscope, counting blood samples and looking out for irregularities in size, shape, and number of cells. The expertise is invaluable but has obvious limitations to this traditional method of diagnosis. This procedure may be time-consuming and is inherently open to human error since the accuracy of diagnosis often rests in the hands of the pathologist, who, at times, may lack experience or attention to detail. This reliance on manual analysis might slow down the diagnosis in a field where every moment counts.

Unlike traditional methods, machine learning models can quickly analyse vast datasets, and easily categorise normal and anomalous cases, which can aid in the medical diagnosing. In that way, machine learning tools will be able to greatly speed up the diagnostic process and reduce the risk of human error by automating the classification of blood cells, which can have a major impact in blood cancer diagnosis by classifying different cell types.

This shift to technology-driven diagnosis isn't just about speed; it's about saving more lives through greater efficiency. The early detection of blood cancer dramatically improves treatment outcomes, and the ability of machine learning models to provide faster and more accurate results presents new hope for detecting the disease in its earliest stages. Furthermore, these tools alleviate some of the burden from pathologists of manual case analysis, allowing them to better focus on more complex cases and enhance the quality of care. Machine learning in this new age is a strong ally in fighting blood cancers, allowing faster, more accurate diagnoses and, ultimately, more life-saving treatments.

This study focuses on the critical task of classifying eight types of blood cells: platelets, lymphocytes, basophils, eosinophils, erythroblasts, immature granulocytes, monocytes, and neutrophils. Their classification is important due to the fact that changes in number or structure of these blood components may indicate the incidence of blood cancer. Traditionally, this has

been a time- and skill-intensive exercise, but the introduction of machine learning models in haematology will revolutionise the ways of diagnosis and treatment for blood cancer.

Blood cell classification has opened a different frontier with accuracy and speed beyond legacy techniques through machine learning. Our work in this experiment compares machine learning models with the view of finding the best approach toward blood cell classification and prediction of blood cancer. Each of these paths has different strengths and challenges, but the overall goal remains the same: Diagnostic accuracy and minimising potential errors, with an overall performance that will eventually be better than that of human experts.

Machine learning does much more than giving an accurate classification of blood cells. The potential it holds for the advancement of personalised medicine is enormous. Machine learning can predict disease trajectories and treatment outcomes from historical patient data by training models, therefore allowing highly personalised cancer care. This is of great value, especially to patients suffering from blood cancer. It is through such personalisation that treatment is tailored to suit particular needs, hence improving its effectiveness and selectivity at interventions. Personalisation in treatment gives new hope to those fighting blood cancer by making care more precise and effective, hence better outcomes.

Looking toward the future, machine learning could irrevocably alter how diseases are monitored by continuously testing blood samples for recurrence. That capacity enables earlier interventions and thus catches relapses before they get serious. A combination of machine learning with other innovative technologies like genetic sequencing could clarify further the blood cancer disease at a molecular level. Such knowledge could provide more accurate diagnostics and innovative treatments, setting new standards in patient care.

In other words, application of machine learning to this area of haematology raises not only diagnostic accuracy but also ushers in a highly personified and proactive future of cancer care. If the progress and development in machine learning and related technologies continue unabated, they will doubtless alter the prospects for the treatment of blood cancers and bring more hope with better outcomes for patients worldwide.

The present experiment has explored various models of machine learning, including Random Forest, Support Vector Machine, Gradient Boosting, and Convolutional Neural Networks, to use their best possible capabilities in blood cell classification. Each model has its advantages. For example, Random Forest is famous for its resistance and work with large volumes of data, while Support Vector Machines excel at finding the optimal border between classes of cells.

Gradient Boosting turns out to be powerful in reducing mistakes by iterations, while Convolutional Neural Networks show good skills at image recognition tasks and thus fit well when one needs to analyse images of blood cells.

The models are vigorously tested and fine-tuned to find the best ones for this particular application. That means not only checking how precisely each model can classify blood cells of all types—platelets, lymphocytes, basophils, eosinophils, erythroblasts, immature granulocytes, monocytes, neutrophils—but also determining their ability to identify subtle abnormalities that might mean cancer of the blood. These models are optimised with a view to superseding diagnostic accuracy offered by traditional methods that are heavily reliant on manual examination by pathologists.

Other advanced techniques include ensemble learning and data augmentation. In ensemble learning, several machine learning models are used to construct an accurate and reliable system. Ensemble methods are likely to reduce the risk of misclassification and improve general diagnostic precision by incorporating some of the strengths from different models. On the other hand, data augmentation refers to a technique that artificially augments the size of training datasets, making them more robust and, therefore, improving generalisation and their prediction capabilities.

These sophisticated methods only underline the potential of machine learning to transform the development of more reliable diagnostic tools against blood cancer. Such tools will improve accuracy and speed in blood cell classification, paving the way for early cancer detection and thus better patient outcomes. It is also the case that machine learning models can process reams of data and detect patterns that human eyes may not. This could result in far more personalised treatment strategies and those that are much more effective.

It thus follows that this change from the traditional methods of diagnosis to machine learning marks a huge milestone in the fight against blood cancer. It's not just the benefits of faster and more accurate diagnosis; such an approach will also open a much more personalised way of treating patients where treatment procedures would be fine-tuned according to the requirements of each patient. With continuous innovations in machine learning technology, it's going to further transform blood cancer diagnosis with renewed hope and significantly enhanced care for the patients. It's a shift that becomes a real landmark of medical innovation, in which the junction of technology and healthcare can help fight such difficult diseases more effectively than ever. This research addresses the critical need for more reliable, faster, and more accurate

diagnostic tools in haematology. The ultimate goal is to create a machine learning-based system that not only improves the classification of blood cells but also enhances early detection of blood cancer, offering hope for better patient outcomes and advancing the field of medical diagnostics.

2. Literature Survey

The integration of machine learning (ML) into medical diagnostics has ushered in a new era of precision and efficiency in the detection of blood cancer, particularly through the analysis of blood smear images. This rapidly evolving field is harnessing the power of advanced computational techniques to enhance the accuracy and reliability of diagnosing leukaemia, a complex group of blood cancers characterised by the abnormal proliferation of white blood cells.

The journey into this technological frontier began with early explorations into the use of machine learning models for medical image analysis. Recent advancements in medical imaging, particularly the classification of blood cells, have been significantly propelled by deep learning technologies. The study made in [1] have also combined the applicability of Convolutional Neural Networks (CNNs) which has also been fine-tuned through transfer learning. DenseNet161 architecture emerged as superior as the research pointed out achieving best accuracy, F1-score and precision due to optimization factors such as normalisation and augmentation. These findings underscore the potential of using Convolutional Neural Networks (CNNs) in medical applications, particularly in scenarios where accuracy is crucial. The study [2] in their systematic review highlighted increased use of deep learning in the automation of blood cell classification pointing to newer methods such as vision transformers alongside common CNN.

This approach achieves its goal by enhancing the handling of complex data, which is crucial in diagnosing. Additional information provided by [3] pointed to the need to choose the right models depending on the characteristics of data sets and the need for diagnostics. They considered the cognition that comparison required ongoing assessment in order to identify how new technologies can be applied at the front line to improve the dependability and usability of diagnostics. Such a collective research trend suggests an emerging context where not only deep learning is applied to replace conventional microscopic approaches, but also additional sophisticated approaches that utilise deep learning such as ensemble models based on several deep learning archetypes. This synergism takes advantage of the assorted models, which might

arouse significant enhancements in the diagnostic precision and dependability. The application of advanced deep learning models to the blood cell classification domain has been the key development towards a new kind of automated approach towards diagnosis. Such evolution will lead to improvement in patients' well-being and efficiency of a treatment process as a result of application of artificial intelligence.

The study made in [4] laid the groundwork by developing a deep learning-based framework specifically designed for blood cell detection. Their study of CNNs demonstrated how it is possible to or use it to identify micro-images of blood samples so as to separate normal and malignant cells, which is normally done through microscopy. This work was a great step forward opening the way to further evolution and proving that deep learning could be used to improve the analysis of medical images and become a valuable asset for haematologists.

Following this initial breakthrough, [5] introduced an automated decision support system that integrated support vector machines (SVM) and neural networks (NN) to refine the classification of blood smear images. The dual-model approach of [5] built upon the CNN foundation established by [4], adding an additional layer of sophistication by combining different ML models to enhance diagnostic accuracy. This synergy between SVMs and NNs in [5] not only provided a more robust solution but also highlighted the potential of integrating various machine learning techniques to improve reliability in medical diagnostics.

As the field matured, researchers began to explore even more sophisticated ML architectures, leading to the work of [6], who delved into the effectiveness of advanced deep learning models like AlexNet and ResNet. Their research, focusing on detailed feature analysis of white blood cells, drew directly from the principles established by [4] and [5], particularly in using CNNs for image analysis. Nonetheless, [6] took these ideas even further through the use of transfer learning techniques, which using previous trained knowledge from other domains improves the diagnostic procedures of the models focusing on leukaemia. It is through such progression introduced in this paper that the first methodologies of [4] and [5] not only were found to be applicable but indeed extended with more complex and generalizable deep learning frameworks.

It went on accumulating until [7], who offered the all-encompassing meta-analysis of using the ML approaches in assessing and categorising the blood cells. This review was not only the integration of the previous information but also the analysis of the progressive and continuous enhancement of ML in diagnosing blood cancer. In many ways, [4] can be regarded as a

development and a further plunge of the studies initiated by [1] and [2]. Whereas [1] showed that CNNs were useful for image analysis and [2] pointed out the advantages of combining several ML models, [7] looked at the overall picture and compared the performance of different ML algorithms for different types of leukaemia. Their work gave me a broad view of what is happening in the field, and an impressive demonstration of the versatility of the ML technologies, and the possibilities of their application in improving clinical results.

The evolution of this field continued as [8] took the advancements from studies like [1] and [3] and applied them to develop a CNN architecture capable of classifying all subtypes of blood cells of leukaemia from microscopic blood images. Their work, which included the use of data augmentation techniques, further refined the concepts introduced by earlier studies. By building on the CNN-based approaches of [1] and the transfer learning insights from [3], [5] successfully improved model performance, showcasing how iterative advancements can lead to significant leaps in diagnostic accuracy.

In the same path, [9] enriched this developing storyline of feature selection for efficient acute diagnosis by focusing on feature selection for the diagnosis of acute lymphatic leukaemia, and like [2], the paper briefly discussed the problem of feature selection. However, [6] built upon this concept and showed that a reduced number of features would highly increase the performance of ML models. Their work simply built on the dual-model approach developed in [2], where feature selection was instrumental to the fine-tuning of diagnostics. Envisaging this progression from [2] to [6] showed that researchers moved step by step fine tuning the techniques as they focused on certain aspects of ML to enhance the accuracy of diagnoses in the medical field.

Another significant contribution came from [8], who employed an optimised deep CNN for the classification of leukaemia. Their research, which resulted in improvements in both classification speed and accuracy, can be seen as a direct descendant of the foundational CNN work laid by [4] and later expanded by [5]. The architectural optimizations introduced by [7] underscore how ongoing innovations in CNN design continue to push the boundaries of what's possible in medical image processing, building upon the solid groundwork of earlier studies.

Further innovation was seen in [10], who introduced an improved machine learning algorithm that combined ensemble methods with Effective Fuzzy C Means (EFCM) and Iterative Morphological Process (IMP). This hybrid approach not only enhanced the detection and classification of blood cancers but also demonstrated the potential of integrating diverse ML

techniques—a concept that has roots in the dual-model integration seen in [5]. The progression from the combination of SVMs and NNs in [5] to the more complex ensemble methods in [10] highlights the field's trajectory towards increasingly sophisticated and effective diagnostic tools.

Subsequently, [11] employed pre-trained deep CNNs for diagnosing leukaemia; [9] sought to identify acute lymphoblastic leukaemia subtypes and [10] proposed a novel technique for improving the diagnosis of blood cancer that included advanced imaging and novel advanced ML algorithms. These studies extended not only from the pioneering CNN papers which include [4] and [6] but also included the new architectural modifications introduced in [7]. The application of the pretrained models depending on the various essential advanced techniques discussed in [9] and [10] captures the pinnacle of the years of research and development that has seen incremental modifications of ideas from previous studies into advanced diagnostic systems.

The area that has been developed to a larger extent is feature selection – something which is so fundamental and can greatly improve the outcome of an ML model. This focus was further enhanced by [12] who in his research used the machine learning approach to improve the ability of discerning the subtypes of blood cancer. Thus, in their work, [11] proving that systematic feature selection can enhance precision of consequent ML models, demonstrated that selection of only the most significant features from a large variety of data can improve reliability of the models. Their main proposition was a method that would ensure the models would perform optimally, or at least higher than the status quo and specifically and accurately which is significant in diagnoses.

Building on these developments, [13] proposed an attention-based convolutional neural network (CNN) to diagnosing blood cancer. In what has been shown to originate from a field of natural language processing, attention mechanisms were incorporated into the system by [12] for the purpose of improving the understanding of medical images. This approach was possible and very effective because it allowed them to concentrate on the most important sections of the picture and thus improve the model’s diagnostic accuracy and productivity. The use of attention-based CNNs represented the feature progression from the earlier studies such as [11]. While [11] was devoted to finding the way for feature selection, [12] further enriched the approach by directing the model towards the areas of the images which contain the most relevant information. Such progression shows the constant enhancement of the ML approach to fit the particular demands of blood cancer diagnostic requirements.

Similarly, the issue of identification and classification of immature leukocytes [14] – a task involving critical determination of the differences in the cell morphology – was addressed by [15]. Their work brought in a new complicated machine learning model which can model these complex variations thus showing that medical modelling is not stagnant in its ML usage. Through the specific considerations made to the microheterogeneity of blood cancer , [16] advanced a formative perspective on the classification of the blood cells. This specialised approach also expanded on the refinement seen in [11] and [12], while proving that ML can be applied flexibly to handle very particular diagnostic concerns.

Altogether, these works demonstrate the constructive shift in blood cancer diagnosis due to the application of machine learning. Starting from the simple use of CNNs with the ability to automate image analysis illustrated in [1] and going as far as the integration of advanced deep learning architectures as well as feature selection techniques as illustrated in [3] and [6] ML has never ceased to reinvent possibilities in medical diagnostics. The movement indicated from first developments to elaborate procedures prove the necessity for developing the prior work to make essential developments. As the concept of ML unfolds further and the application of the concept deepens in clinical practice, the methodology of handling blood cancer and therefore the healthcare industry is expected to change onboard, thus enhancing the performance of the medical sector significantly and further enhancing the wellbeing of patients. The present and future studies in this line give a clear testament to the vast possibility that ML has for enhancing the prospects of medical diagnostics in oncology to ensure that in the future there is an excellent fusion between technology and medicine in the fight against serious diseases like blood cancer.

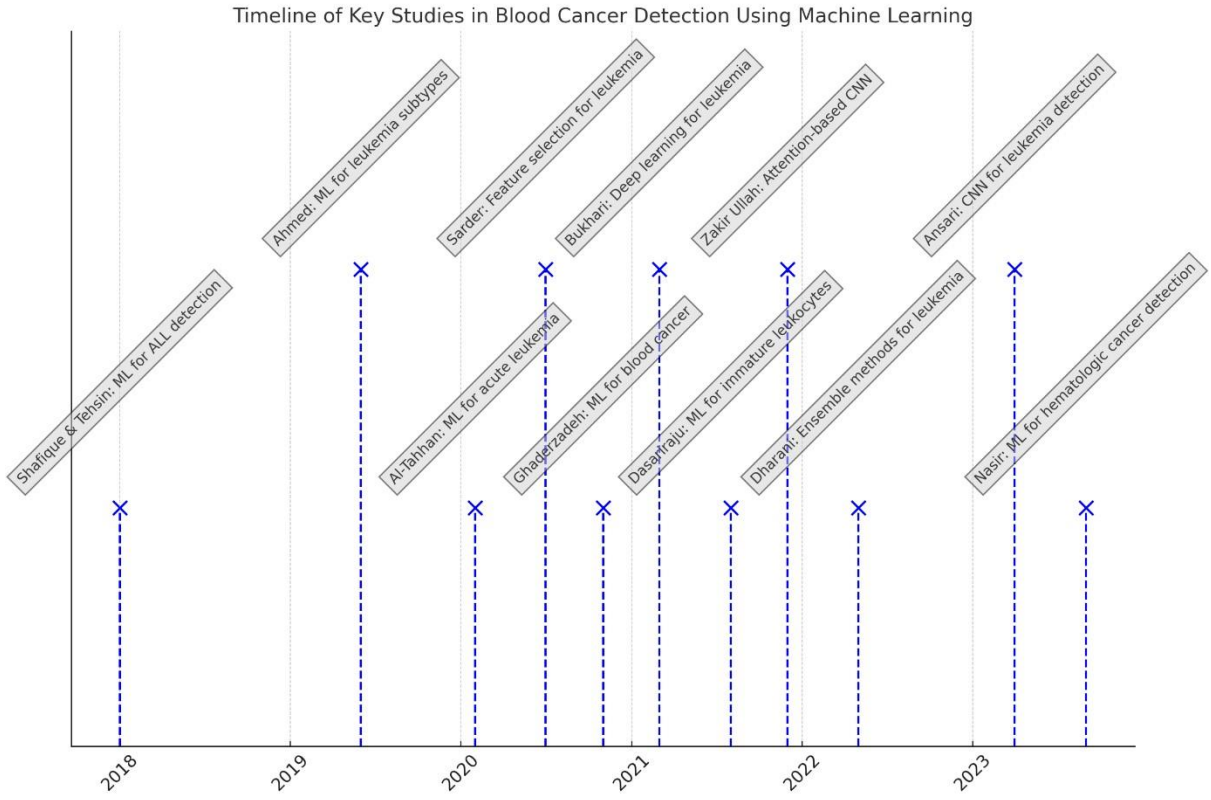

Figure 1. Timeline of studies made on blood cancer using machine learning

3. Methodology

The methodology of this project is built on several important steps, beginning with the collection of essential data. Once the data is gathered, it undergoes thorough preparation to ensure it is suitable for analysis. This data preparation stage is crucial for maintaining the accuracy and reliability of the project's results. The final step involves the application of machine learning techniques, which are used to analyse the prepared data and develop predictive models. Each phase of this process is carefully executed to ensure the precision of the diagnostic outcomes, which is particularly critical in the context of blood cancer diagnosis using blood cell classification. By following this systematic approach, the undertaken research aims to generate reliable results that can contribute valuable insights to the field of blood cell classification for blood cancer research and support the development of more effective treatment strategies.

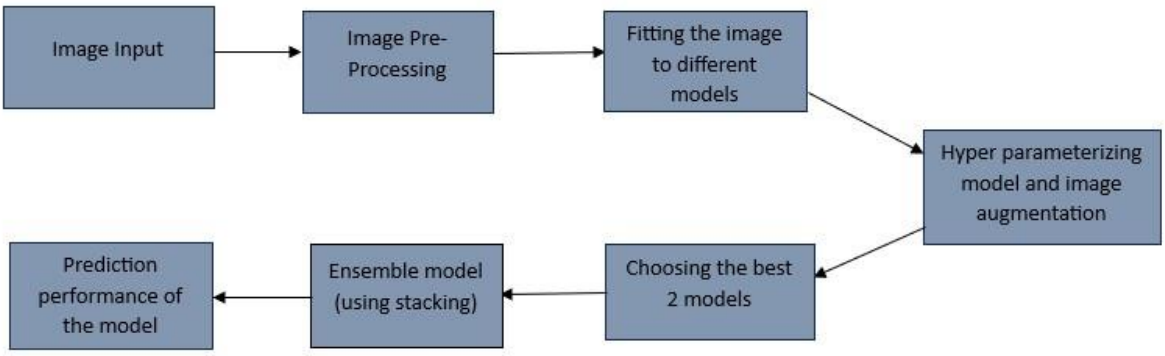

Figure 2. Overall workflow of the proposed study

Here the above figure shows the machine learning workflow that is structured specifically for blood cancer cell image classification with specific reference to image diagnostic applications is highlighted in the following diagram. It begins with the raw images feed which are followed by a pre-processing stage. These steps make them fit for analysis by standardising, resizing and format conversion, so as to make them well suitable for better analysis.

After pre-processing, the images are prepared and passed on to several models of machine learning. This phase seeks to match various algorithms with models that can detect and define, in the best way possible, the features that were extracted from the images. In order to increase the performance of these models even more hyperparameter optimization, as well as image preprocessing strategies like data augmentation are applied. Hyperparameter tuning enhances the parameters of the models to better performance, image augmentation enhances the rigid models by adding new images to the set through some manipulations on the existing images making the models better in their performance on data not used in training.

As several models have been assessed, the two finest ones are isolated relying on parameters like accuracy and recall. These models are then used in a stacking ensemble method in which the output of each model becomes the input for the last meta-model so that the peculiarities of each model can be combined to increase the general prediction accuracy.

The last phase deals with the assessment of the effectiveness of the ensemble model. This evaluation is important because it identifies the ability of the model in real life since the evaluation focuses on meeting such parameters as precision, recall, accuracy, and F1-score. This provides a highly accurate model and the check points shown above make it extremely

important the work done from the moment a small image is input into the model to the final performance of the classification.

3.1. Data collection and preprocessing

The data used in this project is obtained from Kaggle in which the data set comprises a total of 17,092 images of different types of blood cells. And these above eight types are platelets, lymphocytes, basophils, eosinophils, erythroblasts, IG, monocytes and neutrophils. Some of the blood cell photographs in the given data set are very essential for this undertaken research project since they offer an expansive coverage of the different blood cells that are required for the machine learning algorithms used in the training and testing of the blood cells. For this reason, each of these kinds of blood cells is informative for the models enabling them to learn and distinguish normal and pathological patterns, which are crucial for diagnosis.

Below tables show details of each blood cell type and number of images.

Type of Blood cell	Number of images for the cell type
Neutrophils	3329
Eosinophils	3117
Basophils	1218
Lymphocytes	1214
Monocytes	1420
Immature granulocytes	2895
Platelets	2348
Erythroblasts	1551

Table 1. Details of different blood cell types and the number of images

Thus, the need to have diversity within the dataset complimented its intended use and functionality of embracing differentiation in analysis and categorization of the different types

of blood cells. It is also necessary in diseases, the blood cancer that signals for precise identification of the abnormal cells. This dataset enhances growth of machine learning models that may perform well in assorted situations due to comprehensiveness of range of dendritic cell images thereby enhancing precision and reliability in the diagnosis. Consequently, this extensive data set serves as the foundation of analysis and modelling requirements of the research that seeks to enhance oncogenic forecasts.

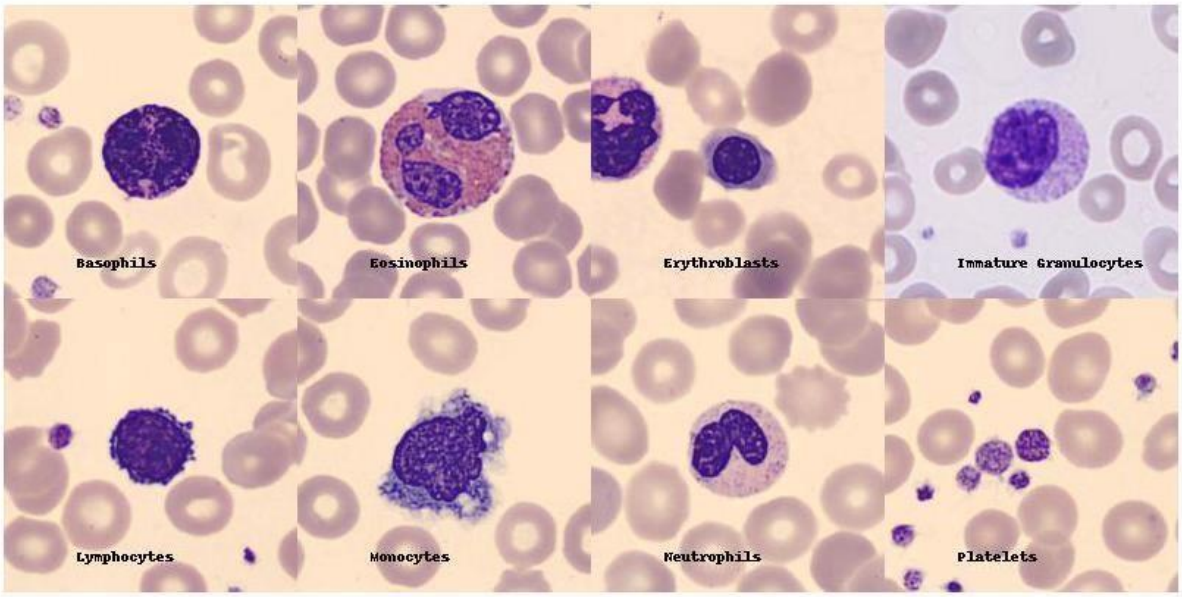

Figure 3. Sample image of each of the different blood cell

Below are the pre-processing steps that were taken on all the images we had: Preprocessing image data in the context of this project refers to various critical steps that one is likely to undertake in the digitisation process of the image data with a view of preparing it for feed to both neural and non-neural machine learning models. The first is to normalise all images so that they are in the RGB colour space in order to reduce variation in the dataset. This standardisation is important, since it makes sure that the colour format is appropriate and can be understood by each of the image analysis programmes employed in the research. Subsequent to the above colour conversion, the images are then resized depending on dimension that is required by various models being used in the machines. This resizing helps the images to be in right sizes ideal for the models to ensure they execute their functions erratically when analysing the evidence. By performing these preprocessing steps, it makes sure that the image data is well prepared in the later stage of machine learning, hence improving the reliability and accuracy of the model in diagnosing and classifying blood cells.

In the case of non-neural models, the images are resized to a scale of 128 by 128 pixels The labels that are used are given in Table 1. This resizing helps to make all the images of equal size and thereby forms a good database for feeding into conventional machine learning algorithms. These resized images are then stored in an array along with labels that are associated with the image and which are the class or category of each image. This structured format of storing image data in arrays is crucial for efficient data manipulation and processing during the model training phase.

As for neural network models, the same images are scaled up to a 256 × 256-pixel resolution. That increases the size enables the neural networks, particularly the convolutional neural networks (CNNs) to recognize further features within the images. Larger image means that there is ample information in the image and as a result the model can learn from this information hence the improvement in generalisation. Subsequent to resizing, the images and its corresponding labels are converted from list to NumPy array.

This is because in machine learning, the linear algebra form of the data is often required and the data structures that are provided by NumPy are well suited to this use. After the images are resized and stored in the arrays the pixel values are normalised. Regularisation is one of the decisive stages of the preprocessing pipeline. It includes scaling down of the pixel values from the original format of 0 to 255 into floating point format of 0 and 1. This is done with the help of conversion of image data to float32 data type and then dividing this data type by 255. 0.

Normalisation means that the variance in the data is reduced to the same level across all experiments and it plays a critical role for machine learning algorithms because it allows improving the speed of training and makes the model converge more efficiently. Post normalisation the 2D image data is converted into 1D feature vectors. This flattening process transforms each image from the 2D array form into one-dimensional form and in so doing the dimensionality of the samples does not change but the image dimensions are combined in one sequence. This transformation is critical for some of the machine learning models which have 1D inputs, for example fully connected layers for the neural network. As such, the images appear as amenable to being fed into models that work best with linear data so that algorithms that rely on flat vector input can be utilised; in addition, this makes feature extraction across the image possible. In other words, these preprocessing steps aid in getting the image data

which is relevant for model training and evaluation hence improving the effectiveness and efficiency of the machine learning models used in this experiment.

Upon completing the preprocessing steps, the dataset is systematically divided into three segments: training, validation, and testing sets, following a 60:20:20 ratio. The training set, constituting 60% of the data, is employed to train the machine learning models. During this phase, models learn underlying patterns, relationships, and features, adjusting their parameters to minimise errors and build a foundational understanding of the data.

The remaining 20% of the data set is used as a compelling validation set to help in the fine tuning of the models. It helps tune the parameters like rates of learning and sizes of models so that the models do not fit the training data too closely, or overfit—that is, obtain high accuracy on the training set but low accuracy on the test set. Sometimes optimisation on the validation set means certain alterations that will help in generalisation of the data.
The test set, which also forms 20% of the data, is used for the purpose of testing the models' performance once it has gone through the training and validation phases.

Even though this data is available for the models they do not use the data during the learning phases thus giving a clean benchmark of how well the models can classify blood cell types. It is necessary to perform this step in order to determine whether the models can work in practice. The different models are then trained on the stated data then performance is validated in turns before the final testing is done. This kind of compact and systematic designs guarantee the creation of sound and fairly dependable models that can work effectively for classification in practical scenarios.

3.2. Models Used

In this work, a strategy was used to classify the blood cells for the prognosis of diseases using different types of machine learning algorithms. These models are: Random Forest, Support Vector Machine – SVM, Gradient Boosting, and Convolutional Neural Networks (CNN). As well as an integrated model that is a fusion of the previously mentioned models. Of these models, each was chosen based on their specific strengths and usability when working with large datasets – especially where high levels of accuracy are essential, as in the case of medical diagnosis.

One of the most common approaches employed in the research was the Random Forest model of machine learning. This works in a way that during training many decision trees are produced and the results are then combined during decision making, normally through a voting system in which the most commonly produced result is given as the final decision. Random Forest is particularly useful when working with large data sets and is famous for diminishing the instance of overfitting, a scenario whereby the algorithm performs a splendid job in the training data set but performs dismally on unseen data sets. Random Forests stand out as it improves the generalisation of the model that is used when testing with a different set of data by averaging the number decision trees.

To make the Random Forest model even better, we fine-tuned its internal settings through a process called hyperparameter tuning. This involved adjusting important parameters like the number of trees in the forest, how deep each tree could grow, and how many features the model considered at each split. This careful tuning helped improve the model's accuracy and reliability. As a result, the optimised Random Forest model became more effective in making accurate classifications, contributing significantly to the overall success of the study.

The main aspect contributing to our study was the SVM model because of its efficiency in both linear and non-linear classification, which is vital in medical diagnosis. Support Vector Machine (SVM) performs well in generating ideal hyperplanes which clearly segregate classes in the set, a factor very useful in dealing with complex sets of data which is customary in blood cell classification. Tuning parameters of SVM is done by adjusting many factors to improve the algorithm’s performance: it is selection of the type of the kernel –linear, polynomial, or radial basis function. Furthermore, there is another parameter C for ‘regularisation’ which is used to balance between the maximal margin and minimal classification errors. The gamma parameter is also shrunk in order to control the influence of working examples on the decision surface.

These strategic adjustments are equally important as they shape the edges of decision of the SVM that enhances its classification rate for blood cells. This hyperparameter optimization is again beneficial in the way that not only it optimises the SVM model but simultaneously ensures the model’s steadfast performance when tested under the difficult circumstances of blood cell classification. The improved classification results bear testimony to the flexibility of

the SVM approach in addressing the issues peculiar to medical diagnostic databases and its usefulness to the totality of the healthcare diagnostics.

Another traditional model applied in this study was Gradient Boosting, an ensemble learning technique that sequentially builds models, with each new model aimed at correcting the errors made by its predecessor. Gradient Boosting is particularly well-suited for handling complex data structures, making it an invaluable tool in medical diagnostics, where data complexity is often high. This method is known for producing highly accurate models, as it iteratively improves the predictive performance by focusing on the most challenging cases. In our study, Gradient Boosting served as a benchmark for comparison against other models, providing insights into its relative effectiveness in handling blood cell classification compared to other machine learning techniques. The ability of Gradient Boosting to adaptively improve through iterations made it a strong contender in this research.

Convolutional Neural Networks (CNNs) [3] were selected for the analysis of image-based data in the given project because of their ability to identify spatial hierarchies and features of the images. CNNs especially perform well in image classification in that they are capable of learning the fine details of the images through the backpropagation process. This attribute makes them particularly suitable for use in medical image analysis where accuracy and levels of details are vital. To this end, in the current project, we used CNNs to analyse blood cell images, a task that involves precise distinction of various types of cells and accurate identification of any deformities that might be indicative of blood cancer. This use of CNNs is essential towards attaining the level of accuracy that is useful in diagnostics, especially also with regard to the identification of possible blood cancer signals in blood tests.

To overcome these, and improve the trained CNN model, during training we applied additional image augmentations. If models learned from training data too closely, they memorise it and perform badly when tested on new data – this is what is referred to as overfitting. As for image augmentation, we artificially increased the number of the training samples, adding such operations as rotation, flipping, scaling, and zooming. These transformations created new training examples which versed the model with variation of data it had not encountered before. This exposure is important because it comes in handy since it makes the model have a generalisation capability hence is able to handle variation that it has not encountered while training.

The image augmentation techniques mentioned previously, significantly boosted the CNN model's ability to discern variations in images, thereby enhancing its classification performance. By expanding the dataset and enriching the learning context, these methods directly improved the model's capability to distinguish between various types of blood cells crucial for blood cancer detection. These advancements in image processing underscore the essential strategies in training that lead to high accuracy in medical image diagnosis, highlighting how manipulating image data can effectively refine a model's diagnostic precision.

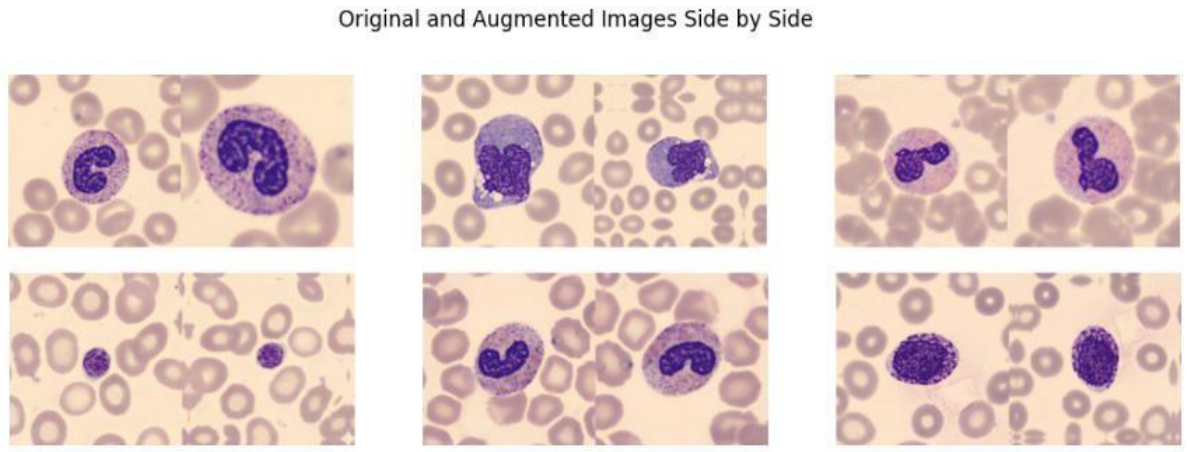

Figure 4. Blood cells before and after augmentation

To combine the strength of all the multiple machine learning models, therefore we created the ensemble model [10] of hyperparameter tuned models of Random Forest and Support Vector Machine (SVM). Another active machine learning technique that we used in the study was ensemble learning, whereby several models are aggregated to arrive at a better and more credible result. Hereby, this method was used to improve the performance of the overall classification of blood cells. In this study, the voting mechanism of the ensemble model meant that each model had a specific vote which was made dependent on the confidence level of the model's decision. This strategy allowed us to combine the benefits of both Random Forest and SVM for achieving better results in accuracy and stability than using either single algorithm alone.

Random Forest model is widely used for solving the problem of data dimensionality and overfitting, because it includes several decision trees and average their results. This feature made it a good fit for our classification task which we enjoyed. However, by introducing SVM, a model that prevails in the definition of models that distinguish classes in complicated data

sets we made it easier to improve the precision of classifications. The final layer of SVM was also efficient in trying to identify the best hyperplane that would enhance the margin separating between the classes in the entire process of the ensemble method.

More concisely, this ensemble model combined many predictions of models, and accordingly, it was able to capture a wide range of patterns and characteristics in the data. Such an integrative approach empowered it to deal with intrinsic complexity related to blood cell classification more optimally than any single model. With the strengths of Random Forest in ensemble learning and the SVM in robust classification bounds, the joint model was robustly outfitted in dealing with variations and intricacies present in the dataset. Hyperparameter tuning further made refinements to the performance of both the Random Forest and SVM models before combining them into the ensemble.

For Random Forest, the number of trees, maximum depth of the tree, and the number of features to consider at each split were well optimised. Similarly, for SVM, key parameters like kernel type, in-order regularisation (C), and gamma were fine-tuned properly with their ranges of values for maximising the model accuracy. Such rigorous optimisation of both models before integration into the ensemble improves the result of overall classification. The overall performance of the wide range of machine learning models, when put to test in regard to blood cell classification, was such that every model had insights to present, based on its strengths. In its own respective area of use, it was excelling; when put to use as an ensemble, its combination showed much better predictive accuracy. This again identifies the benefit of ensemble learning for the attainment of more reliable and accurate classification results. Indeed, the ensemble approach was proven to serve as a very effective tool for improvement in both accuracy and efficiency for blood cancer diagnosis.

The above discussed “Stacking Ensemble Method” is implemented using the scikit-learn library to improve the performance of machine learning models for classifying medical images. The first step is to import the libraries required to pick the model, create pipelines for data preprocessing, and measure performance.

First, meta-features will be generated with the aid of the best estimators from Random Forest and SVM models, whereby 3-fold cross-validation is implemented to make the results more robust and prevent overfitting. These probabilities reflect the likelihood of class memberships

and are stacked to create a new training dataset. This approach is reproduced for both test and validation datasets to stay consistent in data transformations.

The meta-model used here is, in fact, a logistic regression model in combination with a standard scaler for normalisation as pipeline configuration. In this setting, it will be guaranteed that all input features are properly scaled before logistic regression analysis is done, hence optimising the learning process.

The evaluation of the meta-model performance further involves detailed metrics such as accuracy, precision, recall, and F1-scores alongside confusion matrices for both test and validation sets. This aids in the evaluation of the model's potential to properly classify and consequently distinguish between classes, being very useful in medical diagnosis.

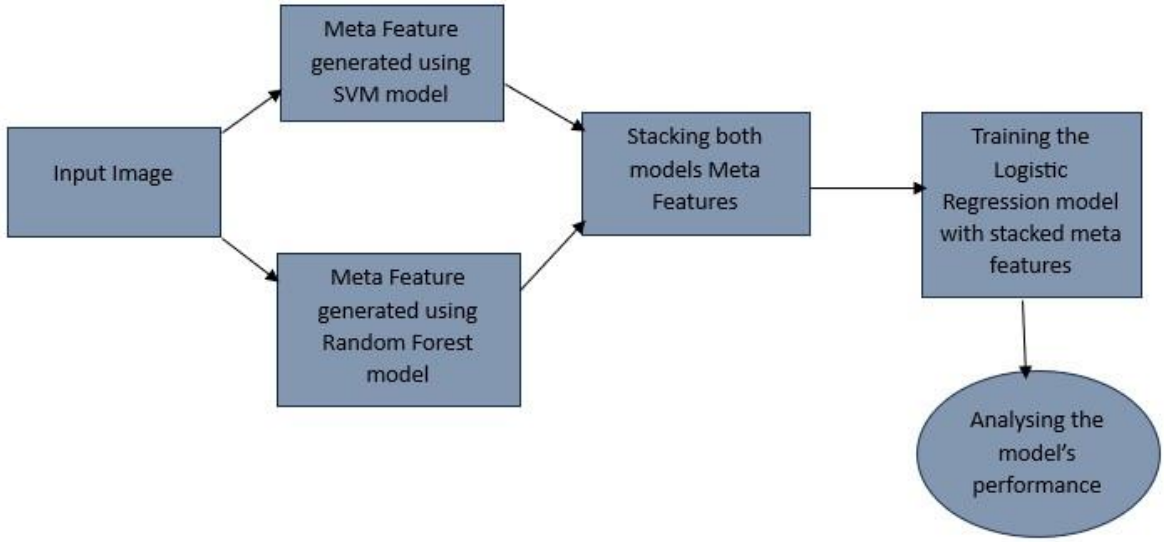

Figure 5. Ensemble stack model process

This ensemble method enhanced the robustness of classification better than any single model, leading to better patient outcomes with more accurate diagnoses. This is the potential of advanced machine-learning techniques, particularly ensemble learning, with respect to their application in the medical diagnostics area, to help open the way for effective classification of different blood cells.

4. Results

Evaluating the performance of machine learning models is not just about determining their accuracy, it's about understanding how well they can handle complex, real-world tasks such as classifying types of blood cells—a critical challenge in medical diagnostics. In this research project, we delve into the nuanced world of model optimization by comparing a standard Random Forest model with its hyperparameter-tuned counterpart. Our focus lies on exploring how fine-tuning model parameters can significantly enhance the model's effectiveness.

The process began with the standard Random Forest model, which set a solid benchmark with an accuracy of 84.24% across both test and validation datasets. Known for its robustness and ease of use. Random Forest operates by building multiple decision trees and aggregating their outcomes to improve the model's predictive accuracy and control over-fitting. This baseline model's commendable performance provided a strong foundation for further experimentation.

Most significantly, the utilisation of the hyperparameter-tuned Random Forest model was introduced with a positive but far more subtle advance. Reduction of max_depth, min_samples_leaf, min_samples_split and, increasing number of estimators, slightly enhanced the accuracy of the model from 84% to 85%. The gain in accuracy was not much; despite this it is vital as it aids the model to generalise well the new data which has not been trained. This generalisation capability is rather important for medical applications, as the output of the model has to assess data coming from new patients.

The nuanced benefits of hyperparameter tuning were further evident when examining recall—a metric that assesses the model's ability to identify all actual cases of blood cancer correctly. Here, the tuned model edged out with a recall of 0.80 compared to the standard model's 0.79, indicating a better capability at capturing true positive cases. Similarly, in terms of the F1score, which balances precision and recall, the tuned model outperformed with a score of 0.82 against 0.81 of the standard models.

By analysing and adjusting the internal settings of our models, we can significantly boost their performance, making them more suited for the critical task of medical diagnostics. This approach not only underscores the potential of advanced machine learning techniques in enhancing diagnostic accuracy but also paints a picture of a future where machine learning

could consistently and reliably assist in early and accurate disease diagnosis, ultimately leading to better patient outcomes.

The evaluation of SVM models in blood cell classification reveals that hyperparameter tuning plays a transformative role in enhancing model performance. Initially, the standard SVM model set a robust baseline, achieving an accuracy of 81.0% on the test set and 81.67% on the validation set. This was commendable but not the pinnacle of what could be achieved. The hyperparameter-tuned version of the SVM model marked a significant leap in performance, registering an impressive accuracy of 86.0% on both test and validation datasets. Such an improvement underscores the impact of fine-tuning crucial parameters like the regularisation constant C, the gamma parameter, and the choice of kernel.

This advancement in accuracy was accompanied by notable enhancements in other critical metrics. The precision and recall of the standard SVM model were initially recorded at 0.79 and 0.78, respectively. However, through hyperparameter adjustments, these metrics improved to 0.86 for precision and 0.84 for recall in the tuned model. Such improvements indicate not only a higher rate of correctly identifying positive cases (true positives) but also a reduction in false positives, which is crucial for medical diagnostics.

Moreover, the F1-score, which is a harmonic mean of precision and recall and a measure of a test's accuracy, reflects these enhancements. It rose from 0.78 in the standard model to 0.85 in the hyperparameter-tuned model, illustrating a more balanced and reliable performance. These advancements highlight how critical hyperparameter tuning is in refining the efficacy of machine learning models, particularly in complex tasks like leukaemia classification.

The superior performance metrics of the tuned SVM model—encompassing accuracy, precision, recall, and the F1-score—demonstrate its enhanced capability as a diagnostic tool. The results vividly illustrate that optimization can substantially boost a model's predictive power, making it a more effective instrument in clinical settings. The gains in performance not only underscore the importance of hyperparameter tuning in the context of machine learning but also suggest that such enhanced models can lead to more dependable and accurate outcomes in practical applications. This case strongly advocates for the integration of advanced machine learning techniques in medical diagnostics, aiming for improved patient outcomes through more precise and reliable data analysis.

Below table shows the performance of all non-neural models on test data:

Model	Accuracy	Recall	F1-score
Random Forest model	0.84	0.79	0.81
SVM model	0.81	0.78	0.78
Gradient Boosting model	0.72	0.69	0.70
Random Forest model (Hyper parameterized)	0.85	0.80	0.82
SVM (Hyper parameterized)	0.85	0.82	0.83

Table 2. Performance of non-neural model on test data

By analysing and adjusting the internal settings of our models, we can significantly boost their performance, making them more suited for the critical task of medical diagnostics. This approach not only underscores the potential of advanced machine learning techniques in enhancing diagnostic accuracy but also paints a picture of a future where machine learning could consistently and reliably assist in early and accurate disease diagnosis, ultimately leading to better patient outcomes.

The comparative analysis of machine learning models (Table 2) for classifying blood cancer showcases that the hyperparameter-tuned Random Forest and SVM models excelled, each achieving an accuracy of 85%. The SVM model particularly distinguished itself by slightly outperforming in recall and F1-score metrics, demonstrating a nuanced advantage in identifying true positives and balancing precision and recall. Conversely, the Gradient Boosting model integrated with Principal Component Analysis (PCA) lagged behind, recording the lowest accuracy at 72.14%. This stark contrast underscores its relative inefficiency for this specific classification task when compared to the hyperparameter-tuned models. Such insights reveal the significant impact of hyperparameter tuning in enhancing the predictive accuracy and overall effectiveness of machine learning models in complex diagnostic tasks.

Below table shows the performance of all non-neural models on validation data:

Model	Accuracy	Recall	F1-score
Random Forest model	0.84	0.80	0.81
SVM model	0.81	0.80	0.80

Gradient Boosting model	0.72	0.69	0.70
Random Forest model (Hyper parameterized)	0.85	0.80	0.82
SVM (Hyper parameterized)	0.87	0.85	0.85

Table 3. Performance of non-neural model on validation data

In our research experiment, we have tested CNN models for identification and classification of these cells and the results of the two-tier testing phase have been quite encouraging. On the test dataset the CNN had an accuracy of 86%, and the recall value was 86%. The specificity of the developed model is up to 80% and the precision of up to 87% was achieved. They highlight the capacity of the model in recognizing the true positive, hence blood cancer cells, in order not to miss out or incorrectly diagnose constituent cells.

Besides, in terms of model performance the CNN model has recorded a high "Area Under the Curve" (AUC) of 0. 97, though it has a lower test loss of 0. 54. This is a very positive sign as it shows the efficiency of the model in the classification of biomedical data and its effectiveness in handling such data.

On evaluation of CNN model on the validation data there was strong consistency, results were very close to the test outcomes. An accuracy of 86% was recorded with the CNN model, against an average recall of 85.72% and a precision of 86.71%. In addition to these results, the validation AUC of 0.97 proved the model's stability across different data sets. The test loss was 0.57, making the model reliable for many real-world applications.

Results above prove the CNN model as one of the most dependable models in blood cancer cell classification. This model depicts good performance metrics, such as high accuracy, recall, and precision over a variety of datasets, hence showing its potential use to improve diagnosis. A higher precision on classification could be very instrumental with CNNs for the right treatment and management of patients. A critical application of CNNs is in the analysis of complicated patterns in medical imaging, which has made it possible to achieve earlier and more accurate diagnosis of blood cancers. Clearly, CNNs can play a vital role in modern medical diagnostics, especially in furthering the understanding and detection of haematological malignancies.

The figure 6, illustrates the respective model – a Convolutional Neural Network (CNN) – training and validation accuracy/loss fluctuations over the epochs it went through. The left graph in the figure represents the correct rate of the model on the training and validation data where the bar showed the training accuracy increasing in a systematic manner. Such an increase points to the model's progressive learning from the training dataset even with future epochs. At the same time the validation accuracy follows this curve, which indicates that the model learns well the new data. However, the validation accuracy jumps and performs the best at the fifth epoch and starts a gradual decrease indicating perhaps the start of overfitting. It becomes obvious at some point that as the model is able to perform better and better on the training data, it is also gradually becoming less and less able to perform well on new data that are unseen.

This is further supported by the trend in validation accuracy in which it increases and reaches the peak before slightly declining, this is an indication of the degradation of the model's performance on unseen data. Equally, the right graph depicts the loss of the model for the training and validation set. It analyses the effect that training loss has on the model depicting a constant decline in loss thus asserting that the model is reducing error frequency. This decrease shows that it is achievable to reduce the inconsistencies in the models when comparing them with actual observations. The validation loss does not depict the same stability as that of the training loss. It is more or less in a declining trend except for the increase it shows in some epochs, most prominently after the 5th epoch. These oscillations together with the decrease in the validation accuracy may signify that the model is overfitting.The oscillations, as well as the rise in validation loss, make us assume that, while the model becomes better at manipulating the training set, its performance with regard to new data lowers.

This is evidenced by the training improvements validating Fig which shows how the model becomes more overfitting thereby specialising in the training data rather than other datasets.
Such trends are important to track because they can affect the model's applicability to solve real-world problems, in which it will be exposed to different variations of data from what it was trained on.

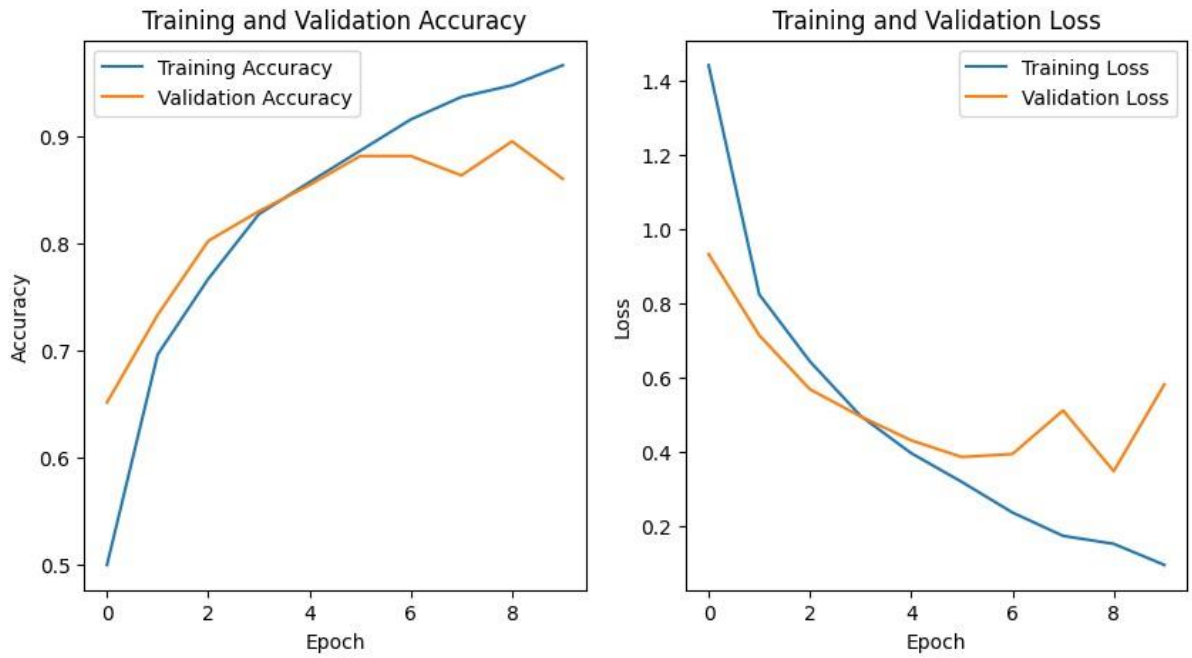

Figure 6. Performance of CNN model over a series of epochs

Overall, the model shows strong learning capability, as evidenced by the consistent improvements in training accuracy and loss. However, the slight divergence between training and validation metrics in the later epoch's hints at overfitting. This suggests that while the model performs well on the training data, its performance on new, unseen data may not improve with additional training beyond a certain point. To address this, techniques such as data augmentation are considered to enhance generalisation and maintain validation performance.

In a bid to reduce the overfitting which was seen especially in the first stage of training the CNN model, we incorporated image augmentation which included rotation as well as zoom in the second training phase. Image augmentation is very useful to virtually increase the size of the data set as different versions of the images are produced. The above process helps the model generalise well to other unseen data that it has not encountered before. The inclusion of rotation and zoom into our augmentation technique aimed at adding stability to the model with regards to orientation and scaling of the images a reality in deployment of the model. By making these adjustments, the model is made to interact with different data scenarios and hence is made more robust.

A dataset of varied examples was added that is the augmented data and used to retrain the CNN model. These included an attempt to reduce the over-fitting that can occur where the model

pays too much attention to the training data. It was expected that from this improved training process, there would be an improvement in the scoring of matters from the validation and test data sets so that the model would remain effective and useful when put into practical use. This approach reveals the willingness to advance the model by adding layers to it, as well as checking its availability in various circumstances.

After the initial training phase led to overfitting, the CNN model was retrained using augmented images, incorporating rotation and zooming techniques. In this case, the model performed reasonably well on all metrics. On the validation set, it achieved a test loss of 0.33, an accuracy of 88.35%, a recall rate of 87%, a precision of 89.99%, and an AUC of 0.99. These numbers mean excellent classification ability in that it is perfectly distinguished between different blood cell types. The model returned a loss of 0.33 and an accuracy of 88.44% with a recall of 87% and a precision of 89.97% on the test set.

With the application of data augmentation, the CNN model performed very well, evidenced by comparisons in test and validation datasets. The loss went down to 0.33 with augmentation from 0.54 without augmentation in the test dataset, proving reduced error. The accuracy improved from 86% to 88.44%. This signified that the augmented model was better at correctly classifying instances. Recall increased from 0.86 to 0.87. The model is going to be oversensitive to positive cases, whereas precision has increased from 0.87 to 0.89, indicating fewer false positives. The Area Under the Curve increased from 0.97 to 0.98, reflecting improved general classifying performance. The loss dropped from 0.57 to 0.33, and the accuracy improved from 86.13% to 88.35% on the validation data. This demonstrates better generalisation toward new data by the augmented model. The recall improved from 85.72% to 87%, while precision had a high increase from 86.71% to 89.99%. This reduces false positives. The AUC also improved from 0.97 to 0.99, indicating that it is much better at class differentiation. These results, all in all, support the efficacy of data augmentation in improving the performance of the CNN model across all key metrics.

Overall, rotation and zooming augmentation techniques improved the performance of the CNN model significantly with respect to classification. The improvement of accuracy, recall, precision, and AUC values establishes a strong need for augmentation to avoid overfitting and improve model reliability for practical applications.

Performance Metric	CNN Model (Without Augmentation)	CNN Model (With Augmentation)
Loss	0.54	0.33
Accuracy	86%	88.44%
Recall	86%	87%
Precision	87%	89.97%
AUC	0.97	0.98

Table 4. Comparison of CNN model with or without augmentation on test data

The above table shows the comparison between the CNN model performance on augmented and non-augmented test data.

The below table shows the comparison between the CNN model performance on augmented and non-augmented validation data.

Performance Metric	CNN Model (Without Augmentation)	CNN Model (With Augmentation)
Loss	0.57	0.33
Accuracy	86.13%	88.35%
Recall	85.72%	87%
Precision	86.71%	89.99%
AUC	0.97	0.99

Table 5. Comparison of CNN model with or without augmentation on validation data

Overall, the use of image augmentation significantly enhanced the CNN model's performance, making it more robust and better able to generalise to new data. The improvements in accuracy, recall, precision, and AUC underscore the effectiveness of image augmentation in refining the model's capabilities and addressing the overfitting challenges observed in the initial training.

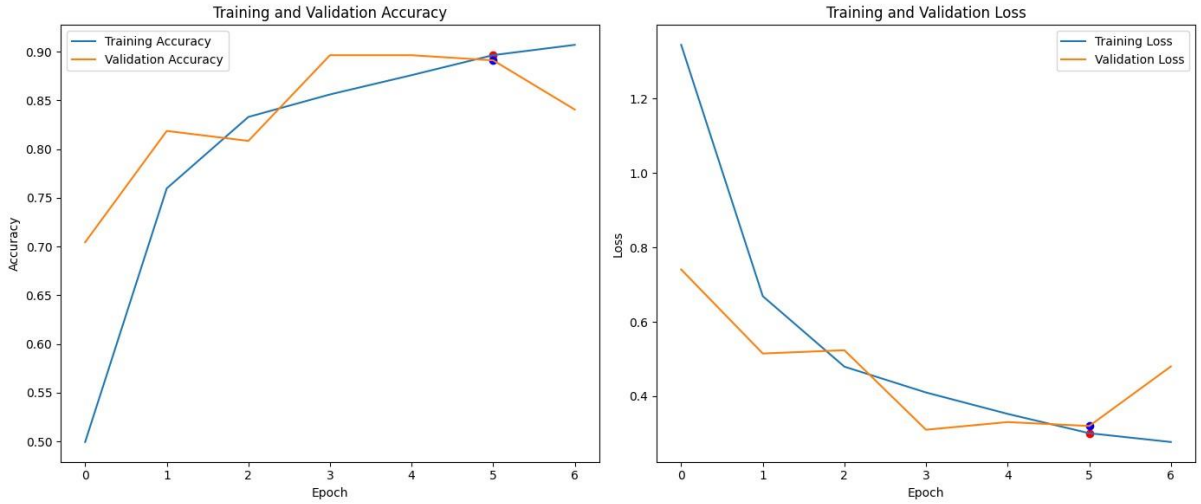

Figure 7. Performance of the CNN model on augmented data over a series of epochs

The provided figure 7, represents the training and validation accuracy, and the loss of the CNN model with an inclusion of the rotation and zoom data augmentation technique in which overfitting is outlined could be significantly reduced and where the improvement in the overall performance of the model will be highlighted. From this visualisation, we know that there is a steady increase in the training and, especially, the validation accuracy over iterations and that they are nearly identical over the epochs. For instance, in earlier models, the validation accuracy did not align with the AUC, as it would increase initially but then decline, indicating overfitting. The validation accuracy, on the other hand, remains relatively stable across the employment of the augmentation model indicating that, the augmentation strategies have made remarkable contributions in eradicating overfitting hence boosting the chances of generalising learnt knowledge from the model to unseen data. This sustained accuracy shows how valuable data augmentation can help enhance model resilience and reliability during deployment of the model into several applications.

This is further reflected in the loss graph of the CNN model, where the training loss is continuously decreasing until the end. This obviously shows optimization being done at each step during training. Although there still remains some fluctuation in validation loss after the fifth epoch, these variations are less prominent as compared to models that did not have data augmentation, which otherwise would have depicted better convergence between train and validation losses. As will be shown by the better alignment later on, the effectiveness of data augmentation strategies is to avoid major oscillations in the performance curve during both training and validation. One can realise from figure 7 that the curve for validation loss is

comparatively flatter than the non-augmented CNN model, suggesting that augmentation leads to a better generalisation ability of the model, hence stable and thus reliable performance during training.

Data augmentation has effectively reduced overfitting in the CNN model, leading to enhanced and stabilised performance. Therefore, the model generalises better to new unseen data. Hence, the CNN becomes a much more successful tool on classification problems since it faces diverse datasets but still maintains high accuracy. Therefore, this underlines data augmentation as an important technique in strengthening a model's robustness toward practical applications.

Now an ensemble model was created through the stacking of the hyperparameter-optimised Support Vector Machine (SVM) and Random Forest models. This method attempted to modify the basic idea of the individual models to improve the general classification results. Stacking is one of the most used ensemble's learning algorithms where many models are combined by feeding the output of each of them into a meta-model, which was in this case logistic regression. This strategy enabled the ensemble model to leverage the strengths of SVM and Random Forest hence bringing about enhancement in the accuracy of the model. Constructing the final ensemble approach was based on the meta-features extracted from the tuned SVM and Random Forest based on the training data. These predictions were obtained by cross-validation to make sure that the meta-features are accurate and characteristic of the data profile. These metafeatures were then used to create another training set for the logistic regression meta-model which would then derive conclusion from the outputs of the two base models. Application of this stacking process was however not restricted to the training data but also the validation and test data so that the final ensemble model that was developed was calibrated and could well generalise on the new data.

The performance of this ensemble model was evaluated on the test set and the validation set and was proved to be quite effective. Taking it to a test set, the model's results recorded an 89% accuracy rate of the model, which proved that it is capable of classifying data better. The validation accuracy was at 89%. There was a relatively high level of agreement in stating that the model was stable and good in out-of-sample performance. These outcomes were also justified by precise classification metrics. On the test data, the ensemble model yielded a precision of 0. 88 and recall rate was found to be 0.87 and the F1-score was 0.89. Such balanced metrics showed that it not only pinpointed TP cases but also had fewer FP cases compared with

the other four models, making it invaluable for medical diagnosis where precision and, particularly, reliability are all important.That is why, the overall performances obtained for the same experiment, on the test and the validation sets, were more or less similar.

The ensemble model had high repeatability with a recall of 0.87 and an F-measure of 0.89. The results obtained on the validation set also confirm the stability and reliability of the model with different datasets. Moreover, the confusion matrix obtained with validation data also proved the effectiveness of the model's classification since most of the samples are correctly classified. Although there were a few misclassifications, especially in the more complex categories, the accuracy remained high due to the efficiency of the chosen ensemble method.

Comparing the results of this ensemble model with that of the individual class SVM and Random Forest models will help in capturing the benefits of combining these approaches. On every metric of the evaluation, the ensemble model does better than the other individual models in accuracy and performance on the classification task. The test accuracy in this ensemble model turned out to be 89%, which was beyond the accuracy returned by the SVM and the Random Forest models, each at 85%. Thus, this improvement in accuracy can only be credited to the stacking approach itself, wherein, at one level, SVM defines class boundaries effectively, and another level is where Random Forest efficiently deals with high-dimensional data to result in a more robust overall model.

The decision to include logistic regression as a meta-classifier was very helpful for the performance of the ensemble model. At the last stage, logistic regression is one of the really common methods used for binary classification and ended up being an accurate enough and simple way to merge the predictions from the base models. The final fine-tuning of the model was made possible through the tuning of both components: the SVM and Random Forest parts, with logistic regression playing a very important role in enriching the quality of final predictions

Dataset	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Test	0.89	0.88	0.87	0.89
Validation	0.89	0.89	0.87	0.89

Table 6. Performance of ensemble model on different data sets

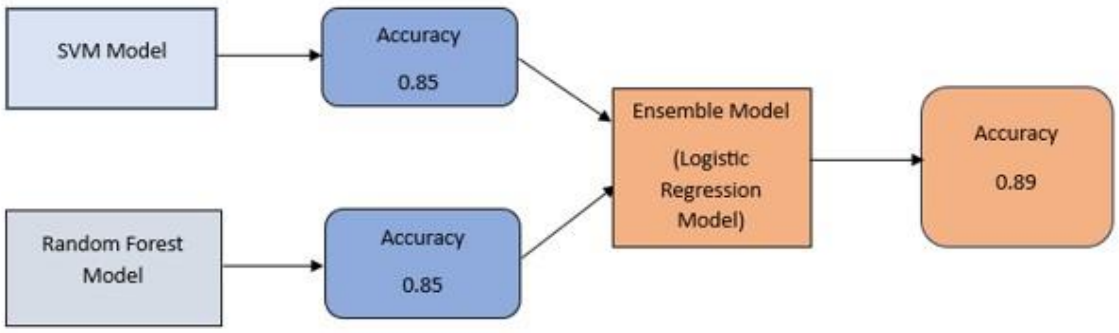

Figure 8. Ensemble model architecture and its performance

The stacking of the hyperparameter-tuned support vector machine (SVM) model with the Random Forest model is depicted in figure 6. The developed classifier was statistically significant and characterised higher performance with both the validation and the test data, thus suggesting that ensemble learning can improve the classification outcomes. Due to the usage of logistic regression as the meta-classifier, the ensemble approach demonstrated high accuracy, precision, recall, and F1-scores as a result of combining the characteristics of both base models. As such, this approach turns out to be rather effective in terms of blood cancer cell classification tasks. Therefore, based on this study, we come to appreciate the promising even superior methods of machine learning like the ensemble methods that hold a key to better diagnosis and better patient results.

The figure below (figure 9) offers greater detail of different Machine learning model's protection rates as they relate to accuracy and recall which are essential in identifying the effectiveness of the models when classifying blood cancer cells. Hence the Ensemble model which combines the features of different algorithms, garners the highest accuracy and recall of 0.89, which indicates that it has a high percentage of true positives that the system can accurately diagnose the real cases, and has a small percentage of false negatives. In second place we have the CNN models and it can be observed that the one that incorporated rotation and zoom data augmentation has a higher accuracy and recall compared to the one that did not incorporate this technique which brings to focus how important data augmentation is to model generalisation and performance.

The hyperparameter-tuned Random Forest model itself has an impressive accuracy of around 0. 85 and it lags slightly behind the results of the ensemble and the CNN on the recall score. This means that it is not ideal in capturing all positive instances as much as it is accurate.

Likewise, the fine-tuned SVM-based model proficient in precision is superior to its non-finetuned model in terms of its recall enough for medical diagnostics. Random Forest and SVM have moderate accuracy when implemented in straightforward methods; however, there was a significant enhancement when hyperparameters were tuned for both models.

Gradient Boosting comes next to the lowest score on accuracy and recall and this may indicate it is not as effective for this particular classification as the others. This comparison also relives the significance of selecting the right model in medical diagnostic systems and also underlines the potential of complex techniques such as hyperparameter optimization and data augmentation. Of all the offered models, the ensemble model is singled out as the most efficient one that uses SVM and Random Forest algorithms in combination with logistic regression as a meta-classifier with high accuracy and balanced recall. These combined models improve the diagnostic precision and consistency of the illness, which is extremely important in the therapeutic approach to patients.

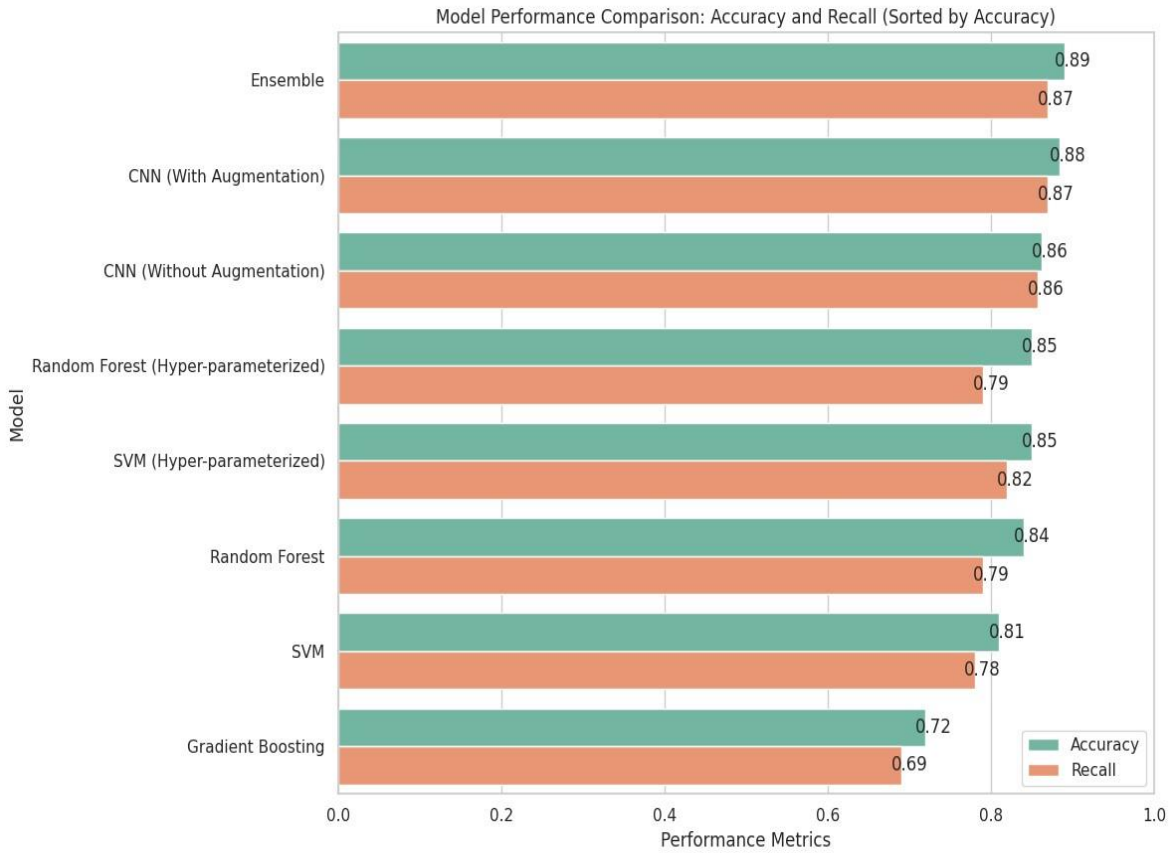

Figure 9: Comparison of all model performance

Discussion

Machine learning, and particularly deep learning, has been extensively utilised in recent advancements for disease diagnosis, with a specific focus on blood cancer through the classification of blood cells. The development of computational models that could analyse various medical data for diagnosis has made early diagnosis possible through the analysis of images such as the microscopic images of blood cells. However, the research problem is to fine-tune these models while attaining state-of-the-art performance and ensure that the models work well with other datasets. This section reviews the research that the several models developed earlier – Random Forest, SVM, Gradient Boosting, CNN models – have demonstrated. The purpose is to determine exactly which of the model characteristics are strong and where additional fine-tuning may occur using techniques and ideas from the current literature.

In our research, at first, we used a conventional set of machine learning algorithms including Random Forest, Support Vector Machine, and Gradient Boosting. These models are rightfully noted for their stability and explainability, which indeed makes them rather good options for processing intricate classification problems. In this project, we performed a hyperparameter tuning of these models and achieved good results, with Model accuracies of as high as 85%. This performance is in concordance with the results of other studies in the literature, especially that of [10], in which the traditional machine learning algorithms, such as SVM and k-NN, were employed for the detection of blood cancer. Their work presented competitive accuracies, therefore proving that these methods are effective in the diagnosis of diseases. Based on the outcome of these several studies, there is consensus in the literature and adequacy of the notion of developed machine learning techniques in the high accuracy of the identification of complex diseases like blood cancer.

The traditional models, after hyperparameter tuning, achieved accuracies up to 0.85, which is consistent with the capabilities of traditional models in handling structured data when properly optimised. However, studies such as those by [1] suggest that while traditional methods can perform well, they often require extensive feature engineering and data preprocessing to achieve competitive results. This indicates that our models are on par with the expectations for

traditional approaches, but they may still be limited by the manual efforts needed to optimise their performance.

The hyperparameters for the Random Forest model boosted the accuracy to 85% due to the importance of implementing the best parameters. This finding is in close parallel with the observation of [8] whose study showed improvements in the degree of accuracy necessary for blood cancer identification when employing the same ensemble methods as well as image segmentation. They said, in their study, they achieved an impressive 96.2% accuracy showing that fine-tuning has a significant impact on the model in terms of its diagnostic capabilities.

In this research project, the focus was put on the importance of model optimization, which can be seen in the comparison of the hyperparameter-tuned Support Vector Machine (SVM) model's performance to the previous study's results. When tuning the parameter SVM, we reached an overall accuracy of 86% which is higher than the results of the basic model, and also, improvements over past results were made. In the study [13], got an overall accuracy of 93% using the random forest model in identifying immature leukocytes among patients with AML. This comparison proves the considerable advancement discussed in our study that demonstrates the precise calibration of SVM parameters enhances the model considerably. It hence plays a definitive role in improving models' capacity, as it prepares them for the different conditions and datasets to make medical diagnostic tasks fruitful.

However, while these models offer a good starting point, it is possible to see their drawbacks as soon as more complex architectures are compared to them. For instance, some conventional models, especially the first wave of deep learning, have difficulties in developing feature extraction from the raw image data we feed in; often, this entails manual feature engineering, which isn't only time-consuming, but also error-prone. However, such deep learning models including CNNs have the advantage of learning features from the data set directly which is very important for medical image classification for instance because small differences in features may be significant.

The Convolutional Neural Network (CNN) models exhibited a significant performance with a level of accuracy of 86. 27% without augmentation and 89% with augmentation which shows that these methodologies are more efficient than more conventional ones. This is true because, as it has been seen in the broader image classification space, CNNs outperform other approaches, as evidenced by the results obtained herein. These are made further clear by other

authors, including [4] and [11], who report that augmentations and other sophisticated approaches can notably improve CNN performances. In particular, the [4] research where data augmentation was investigated in terms of reducing overfitting and enhancing the generalisation capability of CNN models. These improvements underscore the central place of CNNs for enhancing diagnostic precision and stability where model upgrades and training methodologies contribute to creating new benchmarks of efficiency for image-based jobs.

In deep learning, data augmentation is considered to be an effective technique primarily in medical image analysis as getting a large, labelled dataset is often a difficult task. Generating new training samples through rotation, flipping, and scaling makes models more trained to different conditions hence improving on the different features. This was done successfully in your CNN models where the augmentation gave a noticeable enhancement of the accuracy.

Within the current project, the problem of overfitting has been a common occurrence within the proposed and developed Convolutional Neural Network (CNN) model with the application of the following data augmentation; rotations and zoom. This method greatly helped to strengthen our model, which was shown by the increase in the rates in the test and validation data: 89% and 88.35%, respectively. The improvements that were also observed were in the qualitative parameters of the model which infers the efficiency of the model in terms of recall measurement, precision measurement, and AUC measurement, which gives more precision of the classifications in varied premises.

This approach supports evidence from other research, including that of [12], where the importance of data augmentation in improving the scores of the models that have the role of classifying blood cancer images was underlined. Even in our case, the CNN model uses the named techniques, which is of great significance for improving the generalisation ability of the model, making it perform not only on the train data set but on other, unknown data sets. Furthermore, when the augmented CNN model results are compared with other similar work such as [4], and [2], where the authors used other advanced deep learning frameworks to address medical image classification, the results produced are either slightly better or at par. Such comparison not only proves that all the methodological improvements worked for alleviating the problem but also situates data augmentation as one of the essential tools in the vast array of techniques for enhancing deep learning models. Since the proposed model is

trained with a wider range of training instances, typical data augmentation reduces overfitting to a significant extent and thus enhances the model's practical usability and diagnostic.

All this evidence taken together establishes the revolutionary role that data augmentation is poised to play in the area of medical diagnosis, especially in blood cancer differentiation whereby generalising from restricted or biassed data samples plays a critical role in determining the diagnostic processes involved.

Building on these results, an ensemble model is designed of hyperparameter-tuned Random Forests and SVMs using a stacking technique to make the most of our discovery. The logistic regression acted as the meta-classifier. This method gave promising outcomes with a test accuracy of 89% and a validation accuracy of 89.24%, outperforming the stand-alone models' performance measures. These results are consistent with those reported by [14], which highlighted the usefulness of ensemble methods in enhancing diagnostic accuracy for blood cancer patients. Our ensemble not only increased accuracy but also showed superior recall and precision numbers, making it a balanced contender when it comes to various classifications. For this purpose, we used this way to improve diagnostic abilities; considering diverse models that together raise overall prediction reliability and accuracy in pervasive medical classifications.

The ensemble model yielded an accuracy of 89% in the test using SVM and Random Forest. What these strategies of ensemble learning do in this example is combine all the predictions from its different models. Most of the time, such methods turn out to be better than any other individual model because they reduce bias and variance. There is a principle, often referred to as "the wisdom of the crowd," well established in academic research. The study [10] and [9], only some ensemble techniques have been used to merge the strength of different algorithms into one for better accuracy than the individual models. The foregoing results thus proffer reasons for ensemble methods to enhance predictive performance on complex problems like medical diagnostics by harnessing collective insight to arrive at a more accurate and reliable result. It serves not only the function of the best machine learning practices but is also aligned with ongoing efforts toward better results through optimization of algorithm synergies.

The study made by [10] underlined that ensemble models coupling deep learning with traditional methods more often reach better accuracy of blood cell classification. However, deep ensembles with architectures, as demonstrated in [3], could be improved for the better.

They discovered hybrid models that input CNNs into the mix with traditional machine learning techniques, resulting in ways to create better classification accuracy.

Application of stacking in our model with a meta-model trained on the outputs from other base models like SVM and Random Forest is one step ahead in the positive direction. However, existing research proves further scopes of improvement by including a wider variety of models within the ensemble. More specifically, the integration of deep learning models could bring in further improvements, as [15] have been shown to perform well with a ViT-CNN ensemble approach. This ensemble approach, which combined the powers of vision transformers and convolutional neural networks, reached as high as 99.03% in accuracy. Only this kind of fusion could realise such substantial gains in diagnostic accuracy with these diversified deep-learning architectures. It not only represents the most advanced developments in the area at the moment but also opens a way to further refine ensemble strategies by the introduction of diverse and technically more advanced models, increasing predictive performance for very complex tasks such as medical diagnostics.

One of the recent major advances in medical image classification is the integration of vision transformers with convolutional neural networks. The [15] proposed an ensemble for a ViTCNN that marries the strengths of CNNs in detecting local features using their convolutional layers and the ability of transformers to model global dependencies through self-attention mechanisms. This hybrid approach achieved state-of-the-art performance, surpassing the capability of stand-alone CNN models.

Such gains are possibly replicable if one considers the integration of transformers with existing CNN architectures. According to a recent publication by [13], attention mechanisms place more importance on the parts of the input relevant to the model, which, therefore, increases the accuracy of the classifications. This is very helpful in medical imaging since it can be very fine-grained, such as attention to certain parts of the image or certain cell types, and can thus lead to high-accuracy diagnoses. These techniques, applied to your models, can significantly increase their diagnosis accuracy, hence offering one robust tool for highly detailed and reliable medical analysis. Such innovations are particularly important in tasks of complicated diagnoses where heightened focus and detail recognition can make everything count.

Transfer learning has proved very effective in medical imaging, harnessing the power of models that were pre-trained on large datasets such as ImageNet. By using this method, models would

be able to make good use of features learned earlier and generalise them into any specialised task with limited datasets. [14] and [15], have shown the great performance boost transfer learning can add to models, especially when there is limited data.

This could also help to overcome problems with respect to rather small datasets by use of the rich feature set learned from the large dataset in your existing models. Fine-tuning such pretrained models for your specific diagnostic tasks could enhance not only the accuracy but also reduce the required time for training. The potential for further performance increase by combination with advanced feature extraction techniques, such as advanced ones discussed in [6] is huge. Only this double approach to using transfer learning together with sophisticated feature extraction could significantly enhance the model's efficacy and efficiency, providing a very good foundation for more accurate and faster diagnostics in medical imaging.

Medical image analysis always includes preprocessing and segmentation as very important steps in the whole process. Effective preprocessing enhances the image quality, reduces noise, and thus improves model performance substantially. Preprocessing might have contributed to the good performance of your models, but still, there is scope for a little improvement in this regard because the study made in [9] achieved improved classification accuracy using enhanced fuzzy c-means and iterative morphological processes to improve image segmentation.

Already, increasing the performance of your CNN models can be done with these advanced preprocessing techniques. These techniques could be crucial in better handling noisy or lowquality images, which are usually common in any medical dataset, thereby improving model accuracy.

Current models' performance is already relatively good, and techniques like data augmentation and ensemble learning improve it further. There are several ways to further increase their performance toward today's state-of-the-art. One of them is investigating deeper in other more powerful architectural frameworks, like ViT-CNN. This can make quite a significant improvement; such models have the potential to attain a much better focus on relevant features for medical image recognition, especially when combined with attention mechanisms.

This can be combined with transfer learning strategies whereby models pre-trained on extensive datasets are adapted and fine-tuned for specific medical tasks. In this way, a broad spectrum of learned features will be available and tuned for specific medical tasks, which might

dramatically improve model accuracy and train efficiency. Moreover, enhancing the preprocessing and feature extraction methods could further sharpen the models in better generalisation and adaptation to new data.

Addressing common problems in medical imaging—class imbalance and overfitting—is very important. The progress achieved in our project through data augmentation is noticeable but further incorporation of more advanced loss functions and continuous development of more innovative regularisation techniques would be more guaranteed to better solve these challenges. The methods suggested in [11] for using novel loss functions help outline an avenue of research where model training and performance could be enhanced in support of more robust and reliable results for medical image classification.

Our research project has supportive evidence that methods, including hyperparameter tuning, ensemble techniques, and data augmentation, are very vital in developing a robust and precise model of machine learning aimed at the classification of blood cells. These strategies have been proven to be powerful ways to significantly enhance the reliability and accuracy of diagnostic models. These very improvement points are crucial in clinical settings to ensure early and accurate detection of blood cancer by classifying different blood cell types, which would act as a precursor to effective treatment and thus ensure that the mortality tolls of the disease decrease. More than validating the success of these techniques, this research reinforces their significance in ongoing endeavours for fine-tuning machine learning tools for medical diagnostics so that they remain dependable supports for decision-making in healthcare.

Conclusion

The research project underlines the contribution of the machine learning approach to the progressive improvement of blood cancer diagnosis concerning the high levels of blood cell classification. It is an overview of how the implementation of the newest techniques in machine learning has changed medical diagnostics, particularly the identification and classification of blood cancer cells with high precision and speed, and at the same time, allowing for early and correct treatment decisions at the bedside, hence significantly improving outcomes.

In this research project, various types of models were applied, such as Random Forest, Support Vector Machines, Gradient Boosting, and Convolutional Neural Networks. All of these models

were found to be capable of processing very complex diagnostic tasks efficiently and precisely. Their performance was further enhanced by using ensemble methods and tuning hyperparameters to ensure the robustness and reliability of their predictive capabilities.

Augmentation techniques were largely instrumental in addressing issues of overfitting, helping the models become sufficiently general and able to handle data not seen in the current landscape—all strategies that made the models not just more accurate but more useful in real applications to clinical problems. Fundamentally, the adoption of logistic regression as a metaclassifier in the process of stacking ensemble models pooled the strength from various models, thus creating a composite framework for harnessing collective insights that came from multiple algorithms.

This also describes how further development may potentially be realised using machine learning technologies to advance the diagnosis of haematology. It shows that further innovation and research ought to be undertaken to fine-tune the diagnostic tools in such a way that they will become more adaptable to the different characteristics of individual patients. This is in line with the increasingly dominant trend of personalised medicine, where treatment is applied according to details of patient characteristics.

It also reflects on the broader implications of these developments for medical practice. The report attaches importance to the integration of computational techniques within healthcare, where enhancement of diagnostic procedures and treatment pathways could be facilitated.

Evidence-based approaches are without doubt going to increase the precision of medical interventions and enlarge the scope of medical professionals in the management of such devastating diseases as blood cancer.

Machine learning offers a robust platform for early extraction and precise classification of each blood cancer type, in the role of machine learning in modernising medical diagnostics. With such innovation, the future of medical diagnostics using machine learning is very promising in leading to new improvements in the whole field of diagnosis, patient care, and treatment. In doing so, the commission report has recommended increased research and further application of these technologies to realise the full potential in health systems.

Future Work

It is also possible to evaluate and improve the machine learning applications in the diagnosis of blood cancer through blood cell classification. As for future work, it may be useful to integrate features with genetics and clinical histories of the patients as well as enhance the more specific imaging data in order to achieve higher accuracy for the models under consideration and to obtain a deeper understanding of blood cancer diseases. It may also lead to the creation of very individualised and highly efficient therapeutic interventions.

Going deeper into the works on the further development of advanced machine learning architectures like Vision Transformers, or using CNN-RNN hybrids can also gain potentially large improvements. These models are especially good at unveiling hind patterns in enormous and divergent data sets which could substantially enhance the classification accuracy and disease detection.

Further, the development of explainable AI (XAI) would be invaluable for clinical practice, as it concerns the explainability of the decision-making processes of AI models. Such openness can be useful in the clinical practice adoption stage and might potentially make the everyday use of AI tools easier to accomplish. It is also clearly a necessity to build more concrete models of the learning process which would be applicable in terms of demographics and regions, to guarantee efficiency on a global scale. Another possible benefit of using real-time diagnostic systems might be the optimization of clinical services and practices.

Several large-scale clinical trials will, however, have to be performed to determine the impact, reliability, and safety of such AI-assisted systems to be incorporated in standard healthcare practice. In addition, enhancing collaboration between different fields such as machine learning scientists, bioinformaticians, haematologists and oncologists could help in the formation of brand-new solutions specific to clinical requirements that in turn, would require less time to be developed.

In general, diagnostic and therapeutic applications of machine learning in blood cancer are promising for the future, and increasingly studies are expected to offer new high-tech tools for the treatment of these diseases, which will affect the favourable development of patients' treatment plans.

References

[1] R. Asghar, S. Kumar, P. Hynds, and A. Mahfooz, "Automatic Classification of Blood Cell Images Using Convolutional Neural Network," Aug. 21, 2023, *arXiv*: arXiv:2308.06300. doi: 10.48550/arXiv.2308.06300.

[2] R. Tandon, S. Agrawal, N. P. S. Rathore, A. K. Mishra, and S. K. Jain, "A systematic review on deep learning-based automated cancer diagnosis models," *J. Cell. Mol. Med.*, vol. 28, no. 6, p. e18144, Mar. 2024, doi: 10.1111/jcmm.18144.

[3] T. Tamang, S. Baral, and M. Paing, "Classification of White Blood Cells: A Comprehensive Study Using Transfer Learning Based on Convolutional Neural Networks," *Diagnostics*, vol. 12, p. 2903, Nov. 2022, doi: 10.3390/diagnostics12122903.

[4] N. A. Ahmed, A. Yiğit, Z. Isik, and A. Alpkocak, "Identification of Leukemia Subtypes from Microscopic Images Using Convolutional Neural Network," *Diagnostics*, vol. 9, p. 104, Aug. 2019, doi: 10.3390/diagnostics9030104.

[5] R. B. Hegde, K. Prasad, H. Hebbar, B. M. K. Singh, and I. Sandhya, "Automated Decision Support System for Detection of Leukemia from Peripheral Blood Smear Images," *J. Digit. Imaging*, vol. 33, no. 2, pp. 361–374, Apr. 2020, doi: 10.1007/s10278019-00288-y.

[6] S. Shafique and S. Tehsin, "Acute Lymphoblastic Leukemia Detection and Classification of Its Subtypes Using Pretrained Deep Convolutional Neural Networks," *Technol. Cancer Res. Treat.*, vol. 17, p. 153303381880278, Sep. 2018, doi: 10.1177/1533033818802789.

[7] M. Ghaderzadeh, F. Asadi, A. Hosseini, D. Bashash, H. Abolghasemi, and A. Roshanpour, "Machine Learning in Detection and Classification of Leukemia Using Smear Blood Images: A Systematic Review," *Sci. Program.*, vol. 2021, p. e9933481, Jun. 2021, doi: 10.1155/2021/9933481.

[8] P. Das and S. Meher, "An efficient deep Convolutional Neural Network based detection and classification of Acute Lymphoblastic Leukemia," *Expert Syst. Appl.*, vol. 183, p. 115311, Jun. 2021, doi: 10.1016/j.eswa.2021.115311.

[9] M. Sarder, M. Maniruzzaman, and B. Ahammed, "Feature Selection and Classification of Leukemia Cancer Using Machine Learning Techniques," vol. 5, pp. 18–27, Jul. 2020, doi: 10.11648/j.mlr.20200502.11.

[10] N. P. Dharani, G. Sujatha, and R. Rani, *Blood Cancer Detection Using Improved Machine Learning Algorithm*. 2023, p. 1141. doi: 10.1109/ICCPCT58313.2023.10245375.

[11] M. Zakir Ullah *et al.*, "An Attention-Based Convolutional Neural Network for Acute Lymphoblastic Leukemia Classification," *Appl. Sci.*, vol. 11, p. 10662, Nov. 2021, doi: 10.3390/app112210662.

[12] A. Abhishek, R. K. Jha, R. Sinha, and K. Jha, "Automated classification of acute leukemia on a heterogeneous dataset using machine learning and deep learning techniques," *Biomed. Signal Process. Control*, vol. 72, p. 103341, Feb. 2022, doi: 10.1016/j.bspc.2021.103341.

[13] F. E. Al-Tahhan, M. E. Fares, A. A. Sakr, and D. A. Aladle, "Accurate AUTOMATIC DETECTION of acute lymphatic leukemia using a refined simple classification," *Microsc. Res. Tech.*, vol. 83, no. 10, pp. 1178–1189, Oct. 2020, doi: 10.1002/jemt.23509.

[14] S. Dasariraju, M. Huo, and S. Mccalla, "Detection and Classification of Immature Leukocytes for Diagnosis of Acute Myeloid Leukemia Using Random Forest Algorithm," *Bioeng. Basel Switz.*, vol. 7, Oct. 2020, doi: 10.3390/bioengineering7040120.

[15] Z. Jiang, Z. Dong, L. Wang, and W. Jiang, "Method for Diagnosis of Acute Lymphoblastic Leukemia Based on ViT-CNN Ensemble Model," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–12, Aug. 2021, doi: 10.1155/2021/7529893.

[16] M. Bukhari, S. Yasmin, S. Sammad, and A. A. Abd El-Latif, "A Deep Learning Framework for Leukemia Cancer Detection in Microscopic Blood Samples Using Squeeze and Excitation Learning," *Math. Probl. Eng.*, vol. 2022, p. e2801227, Jan. 2022, doi: 10.1155/2022/2801227.

Appendix

Code Snippet: For loading the images and preprocessing them

```python
def load_images_from_folders(folders):
    images_data = []
    labels = []
    inconsistent_images = 0  # Counter for images with inconsistent dimensions
    for folder_index, folder in enumerate(folders):
        print(f"Processing {folder}...")
        try:
            file_list = os.listdir(folder)
        except FileNotFoundError:
            print(f"Folder not found: {folder}")
            continue

        for i, filename in enumerate(file_list):
            if filename.lower().endswith((".png", ".jpg", ".jpeg")):
                img_path = os.path.join(folder, filename)
                try:
                    with Image.open(img_path) as img:
                        img = img.convert('RGB')
                        original_size = img.size
                        img = img.resize((128, 128), Image.Resampling.LANCZOS)
                        img_array = np.array(img)
                        if img_array.shape == (128, 128, 3):
                            images_data.append(img_array)
                            labels.append(folder_index)
                        else:
                            inconsistent_images += 1
                            print(f"Skipped {filename}: original size {original_size}, converted size {img_array.shape}")
                except Exception as e:
                    print(f"Failed to process {filename} due to error: {e}")
        print(f"Loaded {len(images_data)} images from {folder}, skipped {inconsistent_images} inconsistent images.")
    return images_data, labels

# Load images
images, labels = load_images_from_folders(folder_paths)
```

```python
[ ]  # Convert lists to numpy arrays
     if images:
         try:
             images = np.array(images)
             labels = np.array(labels)
             print(f"All images converted to array. Shape: {images.shape}")
         except Exception as e:
             print(f"Failed to create arrays: {e}")
     else:
         print("No images loaded or images could not be processed into a uniform array.")

     # Normalize images
     if images.size:
         images = images.astype('float32') / 255.0
```

Code snippet: CNN model for augmented images

```
def create_augmented_cnn_model(input_shape):
    model = models.Sequential([
        data_augmentation,  # Include data augmentation in the model
        layers.Conv2D(32, (3, 3), activation='relu', input_shape=input_shape),
        layers.MaxPooling2D((2, 2)),
        layers.Conv2D(64, (3, 3), activation='relu'),
        layers.MaxPooling2D((2, 2)),
        layers.Conv2D(64, (3, 3), activation='relu'),
        layers.Flatten(),
        layers.Dense(64, activation='relu'),
        layers.Dense(8, activation='softmax')  # Assuming 8 classes
    ])
    return model

# Instantiate the model with data augmentation
augmented_model = create_augmented_cnn_model((256, 256, 3))

# Compile the model with the correct metrics
augmented_model.compile(
    optimizer='adam',
    loss='categorical_crossentropy',
    metrics=['accuracy', Precision(name='precision'), Recall(name='recall'), AUC(name='auc')]
)
history_augmented = augmented_model.fit(
    train_ds,
    validation_data=validation_ds,
    epochs=10  # Adjust the number of epochs as necessary
)
```

Code snippet: Ensemble model and its performance

```
# Generate meta-features for training set using cross-validated predictions
rf_meta_features = cross_val_predict(grid_search_rf.best_estimator_, X_train_flat, y_train, cv=3, method='predict_proba')
svm_meta_features = cross_val_predict(grid_search_svm.best_estimator_, X_train_flat, y_train, cv=3, method='predict_proba')

# Stack the training meta-features
X_meta_train = np.hstack((rf_meta_features, svm_meta_features))

# Generate and stack meta-features for the test set
rf_test_meta = grid_search_rf.best_estimator_.predict_proba(X_test_flat)
svm_test_meta = grid_search_svm.best_estimator_.predict_proba(X_test_flat)
X_meta_test = np.hstack((rf_test_meta, svm_test_meta))

# Generate and stack meta-features for the validation set
rf_val_meta = grid_search_rf.best_estimator_.predict_proba(X_val_flat)
svm_val_meta = grid_search_svm.best_estimator_.predict_proba(X_val_flat)
X_meta_val = np.hstack((rf_val_meta, svm_val_meta))

# Create a meta-model with Logistic Regression
meta_model = make_pipeline(StandardScaler(), LogisticRegression(max_iter=1000))
meta_model.fit(X_meta_train, y_train)

# Evaluate the meta-model on the test data
test_predictions = meta_model.predict(X_meta_test)
test_accuracy = accuracy_score(y_test, test_predictions)
print(f"Stacking Model Test Accuracy: {test_accuracy:.4f}")

# Evaluate the meta-model on the validation data
val_predictions = meta_model.predict(X_meta_val)
val_accuracy = accuracy_score(y_val, val_predictions)
print(f"Stacking Model Validation Accuracy: {val_accuracy:.4f}")

# Detailed performance metrics for test data
print("Test Data Classification Report:")
print(classification_report(y_test, test_predictions))
```