CHAPTER 3

METHODOLOGY

Amalgamation of image processing and machine learning algorithms with healthcare applications is a popular approach used to implement systems that can help to solve health problems quickly and efficiently. These systems are used as assisting tool by many physicians, especially in fields involving cancer. The use of automated systems is exceptionally useful in the timely identification of chronic diseases like ALL. As an active research topic, ALL-C systems are designed to detect the presence of ALL using microscopic images and is also extended to identify its three stages, L1, L2 and L3. Algorithms that can improve the process of detection and classification are still researched and is considered very challenging as the ALL stages have very slight variations that is difficult to differentiate.

Conventionally, the ALL-C system has two main components, namely, feature extraction and detection. Both these steps are continuously going through revolutions and all these improvements or new algorithms have the common goal of improving the current performance of ALL detection. Developers of ALL should always develop new innovative algorithms in order to meet the continuously changing industrial standards and in order to keep track of latest developments and inventions in image acquisition methods. Both physicians and researchers are constantly search for methods that can improve the working of ALL-C in order to meet the high demand for highly accurate ALL-C system. The developers should also provide solutions that satisfy the above demands and implement systems that are user-friendly, while being accurate. The performance of such systems should be proved individually, before it can be implemented and used by pathologists.

Performance improvement can be achieved in two ways. The first is to use new innovative algorithms Bar-Ilan University (2019) and Greengrad (2019), while the second manner identifies the issues of the existing systems and find solutions to them, which automatically will improve the ALL-C detection performance (Chavolla *et al.*, 2018; Clune, 2019). This research work uses the second strategy and proposes solutions to improve existing algorithms, which cumulatively can improve the performance of ALL-C system.

According to Thang and Pashchenko (2017) and Mosavi *et al.* (2019), the improvement of the working of existing system can be incorporated in two ways. The first identifies issues that are yet to be solved in the existing system and develops solutions to solve them. By solving the issues, the existing system is automatically enhanced. The second manner is to combine two successful algorithms to combine their advantages. This concept is termed as hybridization and has been proved to be successful by several researchers (O'Driscoll *et al.*, 2019; Ren *et al.*, 2019). This research work is based on the second manner of enhancement and proposes algorithms that combine multiple algorithms that can perform a stable, versatile and accurate ALL-C system. The solutions provided in this research work are more compatible for classifying microscopic images as either normal or ALL L1, ALL L2 or ALL L3 and work to improve the accuracy of disease detection.

The accuracy of ALL-C system depends on various factors like microscopic image quality, accuracy of WBC detection and identification of correct set of features to detect ALL. This research work proposes a research methodology that focus on each of these factors separately and enhances them. The algorithms used in the conventional ALL-C system are then replaced by these enhanced algorithms with the hope of improving the overall accuracy of ALL detection. The enhancement operations also include feature selection and improving the classifier used in the last step of ALL-C.

The proposed ALL-C system synergistically combines various algorithms-based on image processing and machine learning algorithms to improve the following steps of ALL-C.

(i) Enhance the quality of the microscopic images,
(ii) To extract the WBCs,
(iii) To construct optimal feature vector using feature extraction and selection algorithms, and
(iv) To enhance the classification step using MLC, DLC and hybrid ML and DC classifiers.

The flowchart in Figure 3.1 presents the various steps used by the proposed ALL-C system that uses the microscopic images for ALL detection and classification.

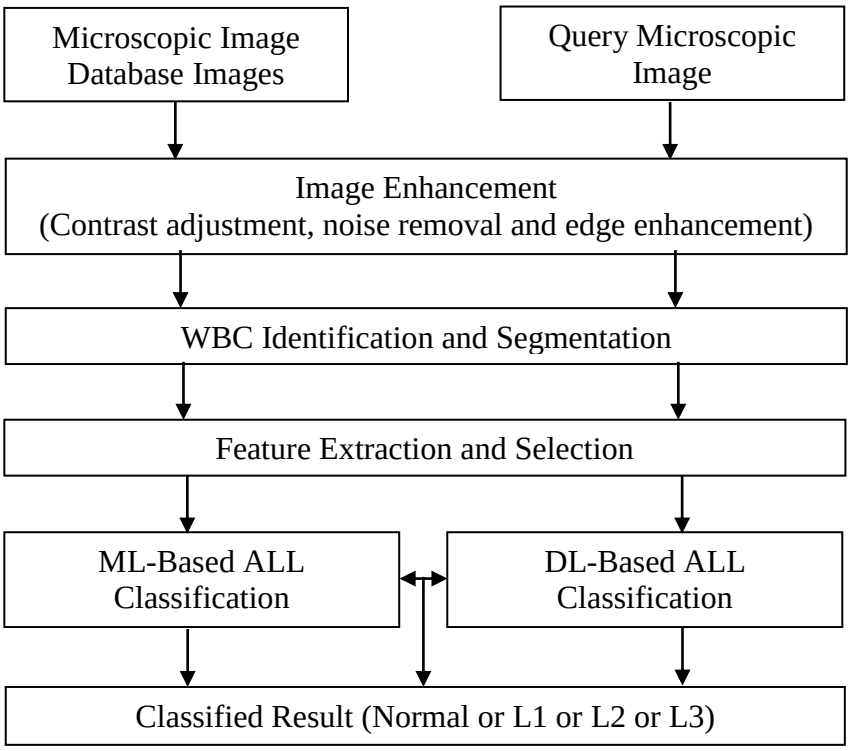

Figure 3.1 : Steps in Proposed ALL-C Systems

3.1. DEVELOPMENT METHODOLOGY, PHASES AND INTERACTIONS

In order to meet the research objectives formulated, the research methodology was designed to have three phases. Each phase is designed with two objectives. The first is to improve the task it is connected to it and the second is to incorporate it with ALL-C system. In general, the aim is to improve the various steps individually, so that its integration improve the overall performance of ALL-C. The integration of phases is implemented using a simple I/O (Input/Output) interface, where the output of the previous phase is used as input by the current phase. Figure 3.2 shows the proposed development methodology.

3.2. PHASE I : PREPROCESSING

The first phase, preprocessing, performs two tasks, namely, enhancement and white blood cell identification. The enhancement task consists of algorithms that improve the visibility of white blood cells in the microscopic images, by handling the degradations caused by the presence of noise, degraded edges and non-uniform contrast. The second task identifies the Region of Interest (ROI) from the enhanced image using a segmentation algorithm. Here, the ROI is the white blood cells in the enhanced microscopic images.

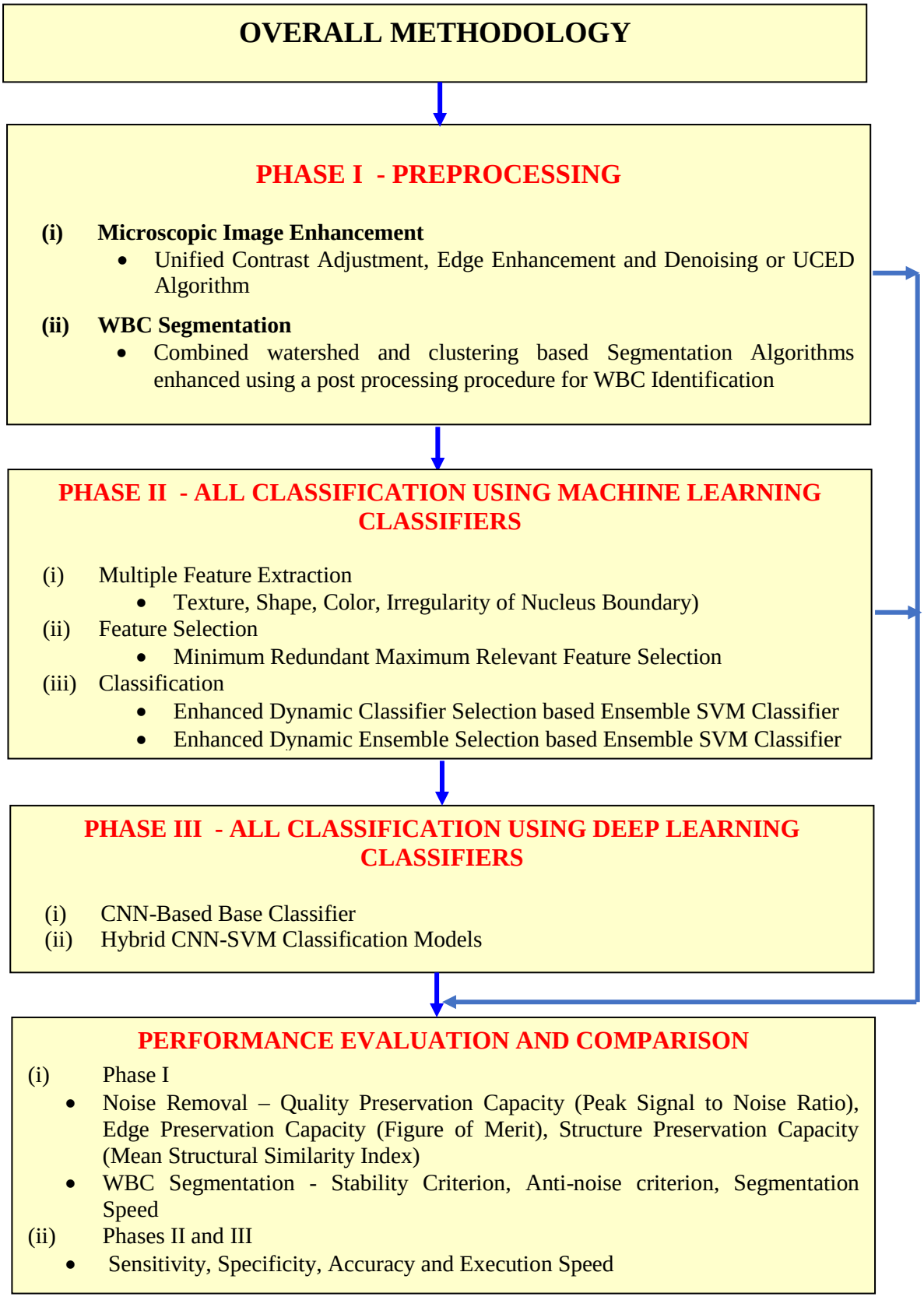

Figure 3.2 : Development Methodology

3.2.1. Noise Removal

The first step of the preprocessing step of ALL-C system is noise removal. The noise in microscopic images is handled by an algorithm that combines the advantages of two frequently used transformation-based algorithms, namely, Discrete Wavelet Transformation (DWT) and K-Singular Value Decomposition (K-SVD). Both the algorithms have their advantages and disadvantages. This research work combines DWT and KSVD to combine its advantages. The proposed noise removal algorithm is an unified approach that combines contrast adjustment algorithm with noise removal and edge enhancement algorithms. The contrast variations are corrected using an adaptive histogram equalization algorithm. The noise present in the microscopic image is removed using a hybrid DWT and K-SVD algorithm. This algorithm beings with DWT coefficients to obtain LL, LH, HL and HH subbands, The LL subband is then divided into edge and non-edge regions using its contrast information. The edge region is enhanced using the sigmoid function, while the noise in the non-edge regions are reduced using K-SVD algorithm.

3.2.2. ROI-Extraction

The second task of preprocessing phase is the extraction of WBC from the enhanced microscopic image. It is considered as one of the most important and challenging task of ALL-C system (Vohra and Prodanov, 2021). A microscopic image has three types of regions, namely, WBC, RBC and background and the goal of segmentation algorithm is to accurately separate these three regions using segmentation algorithm.

In order to extract WBCs, several types of segmentation algorithms are used. Each of these algorithms has its own advantages and disadvantages and when applied to the same microscopic image, produces different segmentation results. Some of these results might be more accurate than others. It is a very challenging job to find the correct segmentation algorithm that can produce stable and accurate segmentation results. The search for the best algorithm that can segment a microscopic image is a very difficult and time-consuming job.

To solve this issue, instead of comparing various algorithms to find a perfect segmentation algorithm, this research work proposes a methodological segmentation design, which attempts to increase the accuracy of segments using two segmentation

algorithms. The motivation behind this methodology is that it is possible to obtain benefits from combining the strengths of multiple segmenting algorithms.

The methodology behind the proposed segmentation method involves two steps. The first step enhances the working of two conventional algorithms, whose results are then combined to form a final set of segments in the second step. The two algorithms considered during the design of the proposed algorithm are the watershed algorithm and K-Means clustering-based algorithm. Both these algorithms were selected because of their success in segmenting images successfully (Vijay and Bhupendra, 2014).

The watershed algorithm is enhanced through the use of a series of techniques, which when applied sequentially can produce accurate segments in a fast manner. The proposed enhanced watershed algorithm is designed using color intensity, Otsu's threshold algorithm, enhanced watershed segmentation algorithm, region merging algorithm and pruning algorithm. The K-Means clustering-based segmentation algorithm is enhanced through the use of an automatic technique to determine the initial seeds using a subtractive clustering algorithm. This research work sets K as 3, since there are three types of blood cells in microscopic images. The algorithm is further enhanced through the use of a computation reduction algorithm, which can speed up the process of clustering and thus, segmentation. The results of the two enhanced segmentation algorithms are then combined by first generating a mean segmentation image, using which a distance map is constructed. Using this distance map, a weight for each algorithm is estimated. Finally, a majority voting algorithm is used to determine the best segment.

The above multiple segmentation-based algorithm is further enhanced through the use of a post processing procedure. The main objective of this procedure is to improve the perceptibility and visuality of the combined clustering results. The operations of the post-processing procedure are listed below.

- Edge Enhancement - As described in Phase I of the study.
- Morphological Dilation - To connect separated points in a better manner using a 2 x 2 structuring element.
- Hole Filling - The internal holes are filled using hole filling method.

Description of the various algorithms proposed in the preprocessing stage are given in Chapter 4, Design of Preprocessing Algorithms.

3.3. PHASE II : CLASSIFICATION USING MACHINE LEARNING ALGORITHM

Phase II of the research methodology uses the segmented results to classify the identified WBC. The steps involved are, feature engineering and classification. Feature engineering consists of two tasks, namely, feature extraction and feature selection. In the classification step, the feature vector obtained from feature engineering is used to classify a cell as normal or cancerous. If cancerous, they are classified into their subtypes L1, L2 and L3.

3.3.1. Feature Engineering

In this research work, multiple features are extracted from the segmented image, as the usage of multiple features, in place of a single feature, help to improve the accuracy of the classifier. Four types of features are extracted (Mostafa *et al.*, 2019). They are, texture features (energy, entropy, contrast, correlation and homogeneity), shape features (area, perimeter, eccentricity, elongation, compactness, minor axis, major axis, solidity, form factor and nucleus-cytoplasm ratio), color features (mean and standard deviation) and irregularity of the nucleus boundary (horizontal direction and vertical direction). Thus, a total of 19 sets of features are extracted.

The above-created set of features consists of irrelevant and redundant features, which induce greater computational cost and may lead to overfitting. To solve these two issues, a feature selection algorithm is used. Feature selection algorithms work on the principle that not all features are important during classification and these features can be removed without affecting the performance. Feature selection is the task of selecting features that have the maximum impact on describing the results and dropping rest of the features with little or no effect on the performance of the classifier. This research work proposes the use of the MRMR (Minimum Redundant Maximum Relevant) feature selection algorithm (Fang *et al.*, 2020) to improve classification performance in terms of both accuracy and speed.

3.3.2. Classification

The final step, classification, is designed as an ensemble classification system. The core idea behind the ensemble classification system is to aggregate a set of learners (known as base classifiers) to obtain a combined classification model, that can produce more accurate results than single classifiers. The basic idea of using ensemble systems is to obtain a more accurate and robust classification by considering multiple views of the same problem. Several researchers have used this idea and proved that multiple classifiers work better than single classifier (Sesmero *et al.*, 2021).

An ensemble classification has three main steps, namely, base classifier generation, ensemble creation and aggregation. Base classifiers in an ensemble system can be created in two manners, namely, homogeneous and heterogeneous. Homogeneous refers to the idea of creating the same type of classifier with different parameters or feature sets, while heterogeneous refers to the idea of using different classifiers trained with the same feature set. In this research work, a homogeneous way of creating base classifiers is used. The classifier selected is the most successful SVM classifier. The bagging subspace feature selection technique is used to create multiple feature sets (100 used in this research work), which are then used to create the ensemble system. The output from each of the base classifier is aggregated using a weighted majority voting algorithm.

In this research work, the above-described ensemble system is enhanced in two manners, as listed below.

(i) Usage of optimization procedure that pre-treat the training feature set, and

(ii) Usage of base classifier selection methods.

The optimization procedure is used to increase the quality of the training set by selecting only those features that help to improve the classification performance. This refined training set helps to obtain a better trained classifier, which in turn helps to improve the accuracy of the ensemble system. The training feature set quality is improved using a two-step process. In the first step, a single SVM classifier is used to pre-treat the training feature set by collecting all the features that produced true positive and false positive results to form a new optimized training set. The features that produced wrong classification results are removed. The newly constructed refined training set can be used

to create a well-trained ensemble system. This step apart from improving the accuracy can also help to reduce the training time complexity of the ensemble system.

The success of the ensemble system depends on the base classifiers selected for classification. Too many base classifiers might increase the time complexity while too few classifiers may decrease the accuracy of classification. Moreover, not all base classifiers constructed are useful during classification. To solve this issue, an additional step, called 'selection' is included. The selection step is used to select only the best performing base classifier(s), which helps to increase the classification accuracy and to reduce the time complexity.

The selection process can be done either using a static or dynamic approach (Yang, 2011). The dynamic approaches can be grouped as (Britto *et al.*, 2014)

- Dynamic Classifier Selection (DCS) - Selects one single best classifier from the set of base classifiers generated, and
- Dynamic Ensemble Selection (DES) - Selects a subset of best classifiers from the set of base classifiers generated.

This research work, to further enhance the process of classification, proposes two enhanced methods that combine static selection and dynamic selection to improve the performance of the proposed EC system. The two proposed hybrid systems are:

(i) EC system using static and dynamic classifier selection method, and
(ii) EC system using static and dynamic ensemble selection method.

The static selection is done using a pruning algorithm that selects optimal classifiers among the base classifiers constructed before the training step. The resultant set of classifiers are then supplied to a dynamic ensemble or dynamic classifier selection method, whose results are used to determine the final classification output.

In this research work, a static pruning technique is used, as a preprocessing function, to reduce optimal candidate classifiers. Static techniques work to construct a subset of base classifiers of fixed size to improve its performance with respect to the full ensemble, removing the rest of the classifiers that do not meet this objective (Munoz *et al.*,

2009). The reason for using a static pruning technique with the proposed enhanced EC system is to produce a smaller-sized base classifier set, which can produce the same advantages of the full ensemble system with added advantages like low time complexity. The pruning algorithm first selects candidate base classifiers whose error rate is low. The kappa statistic pruning algorithm is then applied on these candidate to produce the final set of optimal classifiers.

The aim of dynamic ensemble selection is to find a subset of base classifiers, C_O, that can classify a test sample, such that $C_O \in C_P$ and size (C_O) < size (C_P). In dynamic selection, the classification of test data is done in three steps, namely, region of competence identification, determination of selection criteria and determination of selection mechanism. Region of competence are identified using a K-Means clustering-based method. The selection criterion used is the classification accuracy. The selection method is based either on DCS or DES.

Detailed explanation of the proposed feature engineering and enhanced classifiers are presented in Chapter 5, **ALL-C System using Machine Learning Classifiers.**

3.4. PHASE III : CLASSIFICATION USING DEEP LEARNING ALGORITHM

Currently, there are two methodologies (classical model and deep learning model) available to perform ALL detection and classification. The classical model denotes the usage of hand-crafted features and machine learning algorithms to perform leukemia detection. Developing such a model was the main focus of Phase II of the research work. As ALL detection is a highly sensitive issue related to the health and life of humans, algorithms to further improve ALL detection were probed in Phase III. Phase III, for this purpose, examines the usage of deep the learning algorithm.

To perform early detection of ALL, Phase III proposes a DL classification model based on CNN or ConvNet classifier. This model, called as the base model in this research work, is designed using the AlexNet CNN model based on the transfer learning method, in which deep feature maps were extracted and classified. The CNN classifier was selected because it is one of the most popular types of deep neural networks for image and pattern analysis and recognition (Mayank *et al.*, 2021).

One major issue with CNN is when the size of the training set is small, then the issue of overfitting arises (Xiao *et al.*, 2021). To solve this issue, the training set size is increased through the use of augmentation. Data Augmentation is implemented through the use of image manipulation methods in order to increase the size of the training set. The augmentation methods used in this research work are shifting, rotation, zooming, flipping and shearing. By applying image manipulation methods, the size of the training set was increased and normalized.

The input for the proposed base model AlexNet are the RGB colored images with 227x227 pixel resolution. The base model is designed with 5 convolutional layers with 3 max polling layers. Each convolutional layer is followed by Rectied Linear Unit (ReLU). For transfer learning, the last 3 layers (fully connected layer, softmax layer and classification layer) of the pre-trained AlexNet are used. The proposed AlexNet architecture was modified to include three additional layers, so as to perform classification of 4 classes (Normal, L1, L2 and L3).

In order to solve issues related to time complexity of CNN base model and to further improve the classification accuracy, the CNN base model was boosted through the use of hybrid technology. The CNN base model was modified to include the concepts of the enhanced ensemble SVM model proposed in Phase II. However, the homogeneous manner of ensembling in Phase II was changed to heterogeneous fashion. The ensemble SVM used in Phase III used different kernel type along with the dynamic classifier selection / dynamic ensemble selection methods. Thus, the proposed hybrid model combined the advantages of base CNN model with enhanced heterogeneous ensemble SVM classifier.

Detailed description of the solutions provided in this section is given in Chapter 6, **Classification using Deep Learning Classifiers**.

3.5. EXPERIMENTAL RESULTS

The experiments were performed to evaluate the performance of the proposed algorithms in each phase and study its cumulative effect on ALL classification. The experiments were designed in three stages, with each stage focusing on evaluating the algorithms proposed in them.

Stage 1 focused experiments on the evaluation of the preprocessing algorithms (image enhancement and segmentation). The image enhancement algorithms were evaluated using four performance metrics, namely, peak signal to noise ratio, mean structural similarity index, figure of merit and speed of enhancing a single image in seconds. Visual comparison of segmentation results was used to evaluate the segmentation algorithms. Stages 2 and 3 experiments were used to evaluate the machine learning and deep learning classification algorithms. Four performance metrics, namely, sensitivity, specificity, accuracy and classification speed were used during evaluation. The results obtained from all three stages proved that the enhancement algorithms proposed in each phase are successful and have helped to increase the performance of their respective task. At the same time, the results also proved that all the proposed algorithms have improved the performance of the proposed ALL-C system.

Detailed description of the performance metrics used to evaluate the algorithms proposed in Phases I, II and III along with the experimental results, are tabulated and discussed respectively in Chapter 7, **Results and Discussion**.

3.6. CHAPTER SUMMARY

The development methodology, along with the enhancement plans to improve the various steps of ALL-C, were introduced in this chapter. As mentioned in this chapter, two steps, namely, preprocessing and classification, are used to perform ALL detection and classification. The working of the algorithms proposed for microscopic image enhancement algorithm along with the WBC identification algorithm are described in detail in the following chapter, Chapter 4, **Design of Preprocessing Algorithms.**