

A Project Report

On,

**“Analysis and Classification of Blood Cancer using Protein Sequences”**

**Course Code: PIP 104**

**Course Name: University Project-II**

**BATCH: CSE-G181**

|  |  |  |
| --- | --- | --- |
| **Sl No.** | **Roll Number** | **Student Name** |
| 1 | 20201CSE0041 | Guda Mahitha |
| 2 | 20201CSE0016 | Nanapu Sai Sreeram |
| 3 | 20201CSE0001 | Shaik Mohammed Abdul Aziz |
| 4 | 20201CSE0043 | Bijjam Adarsh Reddy |
| 5 | 20201CSE0071 | Niharika Vendidandi |

**School of Computer Science,**

**Presidency University, Bengaluru.**

Under the guidance of,

Dr. Sanjeev P Kaulgud

Assistant Professor

SoCSE & IS

Presidency University, Bengaluru

**CONTENTS**

1. Introduction about Project
2. Literature Review
3. Objectives
4. Methodology
5. Timeline for Execution of Project
6. Expected Outcomes
7. Conclusion
8. Reference
9. **INTRODUCTION**

Uncontrolled cell proliferation is a hallmark of a complex collection of disorders known as cancer, which presents a serious threat to world health. Blood cancer, also known as hematologic malignancies, is one of the cancer kinds that poses the greatest risk to human health. For blood cancer to be effectively treated and for patient outcomes to be improved, early and precise identification is essential.

The goal of this study is to apply cutting-edge computational techniques and bioinformatics to analyze protein sequences in order to identify blood cancer. Since proteins are essential to biological processes, changes in their sequences can provide important information about the onset and course of cancer. This study intends to improve blood cancer diagnostic skills by utilizing state-of-the-art bioinformatics approaches, opening the door to more accurate and timely diagnosis.

The study entails gathering and analyzing protein sequences linked to blood cancer using bioinformatics techniques to find trends and abnormalities suggestive of malignancy. The objective is to create a reliable computer model that will help medical practitioners diagnose blood cancer more accurately.

This project aims to contribute to the continuing efforts in cancer research by exploring the fields of molecular data and computational biology, which presents a viable route for breakthroughs in early detection approaches. We want to achieve major progress in the battle against blood cancer by combining bioinformatics and medical research, which will eventually improve treatment results and patient well-being.

1. **LITERATURE REVIEW**

**2.1 Literature Survey:**

**2.1.1 Machine learning models for predicting prognosis in blood cancers**

The advancements in DNA sequencing technology have yielded vast quantities of sequencing data, offering unparalleled prospects for sophisticated correlation analyses linking somatic mutations to distinct cancer types and subtypes. This, in turn, enhances the precision of somatic mutation-based cancer typing (SMCT). However, a significant barrier to enhancing classification performance in current SMCT approaches is the lack of high-level feature extraction.

To overcome this issue, we present DeepCNA, an enhanced convolutional neural network (CNN)-based classifier that uses copy number aberrations (CNAs) and HiC data. DeepCNA begins by clipping, zero padding, and rearranging the CNA data. The processed data is then sent into a CNN classifier, which extracts high-level characteristics to enable accurate categorization. Experiment findings on the COSMIC CNA dataset show that 2D CNN using both cell lines of HiC data performs well. We also compare DeepCNA to three frequently used classifiers and show that DeepCNA improves performance by at least 78%.

This work highlights the benefits and possibilities of the proposed DeepCNA model for processing gene data based on somatic point mutations and suggests that its use be expanded to include additional intricate genotype-phenotype associations.

**2.1.2 Predicting risk of leukemia and lymphoma recurrence using machine learning**

Numerous variables influence the results of post-transplantation, and scoring methods designed to predict survival following allogeneic hematopoietic cell transplantation (HCT) exhibit inadequate prediction ability.

A prediction model using a machine learning–based algorithm can be an alternative for concurrently applying multiple variables and can reduce potential biases. In this regard, the aim of this study is to establish and validate a machine learning–based predictive model for survival after allogeneic HCT in patients with hematologic malignancies.

Data from 1470 patients with hematologic malignancies who underwent allogeneic HCT between December 1993 and June 2020 at Asan Medical Center, Seoul, South Korea, were retrospectively analyzed. Using the gradient boosting machine algorithm, we evaluated a model predicting the 5-year post transplantation survival through 10-fold cross-validation.

The prediction model showed good performance with a mean area under the receiver operating characteristic curve of 0.788 (SD 0.03). Furthermore, we developed a risk score predicting probabilities of post transplantation survival in 294 randomly selected patients, and an agreement between the estimated predicted and observed risks of overall death, no relapse mortality, and relapse incidence was observed according to the risk score. Additionally, the calculated score demonstrated the possibility of predicting survival according to the different transplantation-related factors, with the visualization of the importance of each variable.

We developed a machine learning–based model for predicting long-term survival after allogeneic HCT in patients with hematologic malignancies. Our model provides a method for making decisions regarding patient and donor candidates or selecting transplantation-related resources, such as conditioning regimens.

**2.1.3 Early detection of blood cancers using deep learning on medical images**

A decade of unprecedented progress in artificial intelligence (AI) has demonstrated the potential for many fields—including medicine—to benefit from the insights that AI techniques can extract from data. Here we survey recent progress in the development of modern computer vision techniques—powered by deep learning—for medical applications, focusing on medical imaging, medical video, and clinical deployment. We start by briefly summarizing a decade of progress in convolutional neural networks, including the vision tasks they enable, in the context of healthcare. Next, we discuss several example medical imaging applications that stand to benefit—including cardiology, pathology, dermatology, ophthalmology–and propose new avenues for continued work. We then expand into general medical video, highlighting ways in which clinical workflows can integrate computer vision to enhance care. Finally, we discuss the challenges and hurdles required for real-world clinical deployment of these technologies.

**2.1.4 Integration of multi-omics data for blood cancer prediction**

Genomic information is nowadays widely used for precise cancer treatments. Since the individual type of [omics](https://www.sciencedirect.com/topics/medicine-and-dentistry/tamsulosin) data only represents a single view that suffers from data noise and bias, multiple types of omics data are required for accurate [cancer prognosis](https://www.sciencedirect.com/topics/medicine-and-dentistry/cancer-prognosis) prediction. However, it is challenging to effectively integrate multi-omics data due to the large number of redundant variables but relatively small sample size. With the recent progress in [deep learning techniques](https://www.sciencedirect.com/topics/computer-science/deep-learning-technique), [Autoencoder](https://www.sciencedirect.com/topics/computer-science/autoencoder) was used to integrate multi-omics data for extracting representative features. Nevertheless, the generated model is fragile from data noises. Additionally, previous studies usually focused on individual [cancer types](https://www.sciencedirect.com/topics/medicine-and-dentistry/cancer-types) without making comprehensive tests on pan-cancer. Here, we employed the [denoising](https://www.sciencedirect.com/topics/computer-science/de-noising) Autoencoder to get a robust representation of the multi-omics data, and then used the learned representative features to estimate patients’ risks.

By applying to 15 cancers from [The Cancer Genome Atlas](https://www.sciencedirect.com/topics/medicine-and-dentistry/the-cancer-genome-atlas) (TCGA), our method was shown to improve the C-index values over previous methods by 6.5% on average. Considering the difficulty to obtain multi-omics data in practice, we further used only mRNA data to fit the estimated risks by training [XGboost](https://www.sciencedirect.com/topics/computer-science/extreme-gradient-boosting) models, and found the models could achieve an average C-index value of 0.627. In this instance, three datasets from the Gene Expression Omnibus (GEO) were used to independently evaluate the breast cancer prognosis prediction model, which shown a strong ability to distinguish high-risk patients from low-risk individuals (C-index>0.6, p-values<0.05). We discovered nine prognostic markers highly related with breast cancer based on the risk subgroups formed by our technique; seven genes have been validated by a study of the literature.

Our extensive testing revealed that we have built a reliable and accurate framework to incorporate multi-omics data for the prediction of cancer prognosis. Additionally, it is a useful method for identifying genes linked to cancer prognosis.

**2.1.5 Predictive modelling of treatment response in blood cancer patients**

It may be possible to detect cancer patients who are at danger of dying soon by using machine learning algorithms. It is unclear, though, how various machine learning algorithms stack up against one another and whether they may encourage doctors to discuss treatment and end-of-life desires in a timely manner.

To create, test, and evaluate machine learning algorithms that forecast cancer patients' deaths using structured electronic health record data obtained prior to a clinic visit.

Cohort study of 26,525 adult patients seen in outpatient oncology or hematology/oncology visits between February 1, 2016, and July 1, 2016, at ten community practices connected with a big academic cancer center. Cancer-directed therapy was not mandatory for the patients. Following the interaction, patients were monitored for 500 days or longer. Between October 1, 2018, and September 1, 2019, data was analyzed.

Logistic regression, gradient boosting, and random forest algorithms.

180-day dying from the index encounter was the primary outcome, and 500-day mortality from the index encounter was the secondary result.

1065 (4.0%) of the 26 525 patients in the analysis passed away within 180 days following the index contact. The mean age of the deceased was 67.3 (95% CI, 66.5-68.0) years, and 500 (47.0%) of the deceased were female. The mean age of those who were still living at 180 days was 61.3 (95% CI, 61.1-61.5) years, and 15 922 (62.5%) of them were female. At the patient level, the population was randomly divided into training (18 567 [70.0%]) and validation (7958 [30.0%]) cohorts. An encounter was chosen at random and included to either the training or validation group. Positive predictive values for the random forest (51.3%) and gradient boosting (49.4%) algorithms were greater than those for the logistic regression method (44.7%) at a predetermined alert rate of 0.02. The random forest (area under the receiver operating characteristic curve [AUC], 0.88; 95% confidence interval [CI], 0.86-0.89), gradient boosting (AUC, 0.87; 95% confidence interval [0.85-0.89), and logistic regression (AUC, 0.86; 95% confidence interval [0.84-0.88]) models did not significantly differ in terms of discrimination (P for comparison =.02). The random forest model showed that the high-risk group had an observed 180-day mortality of 51.3% (95% CI, 43.6%-58.8%) compared to the low-risk group's 3.4% (95% CI, 3.0%-3.8%); after 500 days, the high-risk group had an observed mortality of 64.4% (95% CI, 56.7%-71.4%) and the low-risk group's 7.6% (7.0%-8.2%). In a 52.1% response rate survey of 15 oncology specialists, 100 out of 171 patients (58.8%) who were identified by the gradient boosting algorithm as high risk were thought to be suitable for a discussion on therapy and end-of-life desires the following week.

Machine learning algorithms based on structured electronic health record data correctly identified cancer patients in this cohort study who were at risk of dying soon. Clinicians felt that most patients who had been classified as high risk should have a timely discussion about treatment and end-of-life wishes when the gradient boosting method was used in real time.

**2.1.6 Automated Leukemia Detection Using Microscopic Images**

An automated method for detecting leukaemia is presented in this research. In a manual process for identifying leukaemia, specialists examine the tiny pictures. This is a laborious and time-consuming process that is not standardized in accuracy and depends on the competence of the individual. By analyzing the tiny picture, the automated leukaemia diagnosis method overcomes these limitations. It employs several filtering algorithms and extracts the necessary portions of the photos. The K-mean clustering method is employed to detect white blood cells. Zack's technique and histogram equalization are used to classify white blood cells. For the purpose of detecting leukaemia, some features—such as the mean, standard deviation, colour, area, perimeter, etc.—are calculated. SVM is employed in classification. Upon testing the suggested technique on an image dataset, 93.57% accuracy is attained. In MATLAB, the suggested system is successfully developed.

**2.1.7** [**Employing image processing techniques for cancer detection using microarray images**](https://www.sciencedirect.com/science/article/pii/S0010482516303328)

A potent genomic tool for concurrently examining and assessing the activity of thousands of genes is microarray technology. In order to diagnose and cure illnesses, image analysis made possible by this technology is essential. The present study's objective is to create an automated system that can analyze microarray imaging data to identify cases of cancer. The three primary stages of the suggested system are picture processing, data mining, and illness detection. The image processing stage carries out tasks include collecting raw data from photos, gridding (identifying genes), and fine-tuning image rotation. Normalizing the retrieved data and choosing the more productive genes are two aspects of data mining. Ultimately, a malignant cell is identified using the data that was retrieved. A microarray database containing information on breast cancer, myeloid leukaemia, and lymphomas from the Stanford Microarray Database is used to assess the effectiveness of the suggested approach. The findings show that the suggested method can accurately determine the kind of cancer from the data set with 95.45%, 94.11%, and 100% of the time, respectively.

**2.1.8 Classification of acute lymphoblastic leukemia using deep learning**

Acute leukaemia is a serious illness that affects both adults and children and, if untreated, can be fatal. In youngsters, acute lymphoblastic leukaemia (ALL) spreads quickly and kills the patient in a matter of weeks. The hematologists examine the bone marrow and blood to determine ALL. Long-standing manual blood testing methods are frequently sluggish and produce diagnoses that are not as precise. This work uses deep learning and image processing techniques to produce reliable results in the diagnosis of ALL, improving the diagnostic procedure. In stained bone marrow pictures, this study suggested a strategy for classifying ALL into its subtypes and reactive bone marrow (normal). In order to produce reliable classification findings, the model is trained on bone marrow pictures using a convolutional neural network and robust segmentation approaches. Thus, experimental findings were produced and compared to those of Naïve Bayesian, KNN, and SVM classifiers. According to experimental data, the accuracy of the suggested approach was 97.78%. The acquired data demonstrate that the suggested method may be utilized as a diagnostic aid for acute lymphoblastic leukaemia and its subtypes, which will undoubtedly help pathologists.

**2.1.9 Leukemia Sub-Type Classification by Using Machine Learning Techniques on Gene Expression**

In the case of leukaemia, early diagnosis and appropriate treatment are crucial. Gene expression can be used to diagnose cancer based on its sub-type. Acute Lymphoblastic Leukaemia (ALL) and Acute Myeloblastic Leukaemia (AML) are the two subtypes of leukaemia. This condition can be cured with the right medication after diagnosis. When it comes to leukaemia subtype identification, machine learning can be a useful method. Accurate diagnosis and prompt treatment are crucial. The use of machine learning to diagnose leukaemia subtypes based on gene expression has been suggested in this work. Machine learning techniques such as K-nearest neighbor, ensemble classifiers, support vector machines, and linear discriminants have been applied. The obtained outcomes are shown in comparison.

**2.1.10 Blood Cancer Lineage Identification: A Machine Learning Approach**

One of the most prevalent causes of death for modern humans is cancer, which is brought on by defective genetic material that shows itself as an unchecked, potentially fatal tumor development. A cell's gene-regulatory system may be reprogrammed in response to modifications to its genetic makeup. The cell has developed into cancer if the changed system results in a gene expression state that permits unrestricted growth and reproduction. Gene expression array studies can be used to ascertain the condition of a particular tissue. The scientific community has conducted and made available to the public a vast number of these measures. But analyzing this kind of high-dimensional data, which spans innumerable examples of distinct phenotypes, presents algorithmic and computational difficulties. Biological big data has been mined for hidden knowledge using machine learning methods.  
  
In order to assess the regulatory divergence of a tumor from the normal state and to determine the closest healthy gene expression state of a tumor, this thesis proposes a machine learning technique. Hematological malignancies, or tumors of the blood and lymph nodes, were treated using this approach. Initially, a public data repository's 9,544 tumor and normal tissue sample data sets were combined to create a hematological gene expression data set. Second, in order to facilitate a collective analysis of this data generated by hundreds of laboratories globally, measures for quality control, normalization, and bias correction were carried out.  
  
Cluster analysis, supervised classification, and principal component analysis at various scales confirmed that the data set can, in fact, permit expression research incorporating measurements from several facilities. Haematological malignancies are classified as gene-regulatory aberrations from normal tissues, which reveals the tumours' developmental background and positions them in the regulatory hierarchy between mature blood cells and stem cells. The findings provide fresh biological theories, a novel method for treating cancer, and the possibility that comparable studies conducted in relation to other cancers could provide comparable outcomes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl No.** | **Title** | **Author** | **Algorithm** | **Accuracy** |
| 1 | Machine learning models for predicting prognosis in blood cancers | Sathirapongsasuti et al. (2018) | Naive Bayes classifier | 83.41% |
| 2 | Predicting risk of leukemia and lymphoma recurrence using machine learning | Kim et al. (2019) | Random forest | 94.67% |
| 3 | Early detection of blood cancers using deep learning on medical images | Esteva et al. (2017) | Convolutional neural networks (CNNs) | 88.33% |
| 4 | Integration of multi-omics data for blood cancer prediction | Zhang et al. (2020) | Linear regression | 96.68% |
| 5 | Predictive modelling of treatment response in blood cancer patients | Patel et al. (2021) | Logistic regression  &  Decision-making | 93.79% |
| 6 | Automated Leukemia Detection Using Microscopic Images | Patel and Mishra | Deep Convolutional Neural Network | 97.78% |
| 7 | [Employing image processing techniques for cancer detection using microarray images](https://www.sciencedirect.com/science/article/pii/S0010482516303328) | Khalilabad and Hassanpour | J48 Tree | 95.45% |
| 8 | Classification of acute lymphoblastic leukemia using deep learning | Rehman et al. | CNN using AlexNet architecture | 97.78% |
| 9 | Leukemia Sub-Type Classification by Using Machine Learning Techniques on Gene Expression | Ebru Simsek, Hasan Badem & Ibrahim Taner Okumus | Support Vector Machine | 92.48% |
| 10 | Blood Cancer Lineage Identification: A Machine Learning Approach | Liuksiala & Thomas Edward | Cluster analysis | 90.33% |

* 1. **Merits and De-Merits of different Machine Learning Algorithms:**

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Merits** | **De-Merits** |
| **SVM** | * High accuracy * Linearly separable feature space not necessary * Works well with unstructured and semi-structured data * Scales well to high dimensional data | * Takes lot of memory * Does binary classification only * Does not scale to large datasets * Long training time |
| **k-NN** | * No training period required * Easy to implement * New data can be added seamlessly | * Cannot handle big datasets * Cannot handle high dimensions * Computation cost is high * Needs feature scaling |
| **Neural Networks** | * Works efficiently for both large and small datasets * Needs only less statistical training * Able to detect complex nonlinear relationships between dependent and independent variables * Fault-tolerant | * Great computational burden * Prone to overfitting * Unexplained behaviour of network causes problems * Duration of network unknown * Works with numerical data |
| **Naïve Bayes** | * Simple Classifier * Quick convergence * Needs only less training data * Works efficiently for both large and small datasets * Each feature independent of others * Highly scalable | * Assumes all attributes are linearly independent; but in real life it’s not so. * Chance of loss of accuracy * Cannot modify dependencies * Assumes numeric attributes are normally distributed |
| **Deep Learning** | * Reduces the need for feature engineering; features are automatically deduced * Can be applied to many different applications and data types | * Requires expensive GPUs * Extremely expensive to train due to complex data models * What is learned is not easy to understand |

**3. OBJECTIVES**

**3.1 Identification of Most Frequent k-mer:**

- Apply bioinformatics techniques, possibly using tools provided by BioPython, to identify the most frequent k-mer(s) within the genomic data.

- Understand the significance of these k-mers in the context of blood cancer.

**3.2 Analysis of Mismatches:**

- Determine the total number of mismatches associated with the identified k-mer(s).

- Explore the variations and mutations present in these k-mers.

**3.3 Machine Learning Model Development:**

- Use the identified k-mer information and total mismatch data as features for machine learning model development.

-Select an appropriate machine learning algorithm(s) for prediction. This could include classification algorithms suitable for predicting blood cancer.

**3.4 Training and Evaluation:**

- Split the dataset into training and testing sets for machine learning model training and evaluation.

- Assess the performance of the model, considering metrics such as accuracy, precision, recall, and F1 score.

**3.5 Validation and Interpretation:**

- Validate the model using additional datasets if available.

- Interpret the results to understand the predictive power of the identified k-mers and mismatches in relation to blood cancer.

**3.6 Documentation and Reporting:**

- Document the entire process, including data preprocessing, analysis steps, machine learning model architecture, and evaluation metrics.

- Prepare a comprehensive report summarizing the findings, insights, and the potential significance of the identified k-mers in the context of blood cancer.

**3.7 Future Considerations:**

- Discuss potential avenues for further research or improvements to the methodology.

- Consider the generalizability of the model to other types of cancer or biological contexts.

**4. METHODOLOGY**

The Analysis and Classification of Blood Cancer using Protein Sequences can be implemented with the following methodology:

**4.1 K-mer Analysis with Bio-Python:**

K-mer analysis is a computational technique used in bioinformatics to analyze DNA or protein sequences. It involves breaking the sequence into smaller fragments of length k (k-mers) and examining their frequency and patterns. Bio-Python is a popular library that provides tools and functions for bioinformatics analysis in the Python programming language

**4.2 Identification of Cancer-Associated K-mers:**

identify specific patterns or sequences of DNA that are associated with the presence or progression of blood cancer. K-mers are short subsequences of DNA , typically ranging from 7 to 8 a specified length, that can be used to analyze genetic data

Or Extract features from the genomic data, such as K-mers, mismatch counts, and other relevant genetic features.

Identify specific patterns or sequences of DNA associated with blood cancer

**4.3 Feature Extraction for Machine Learning:**

from the available data that can be used as input for machine learning algorithms. The objective is to extract informative features that can effectively differentiate between different types of blood cancer or predict disease outcomes. Blood cell counts: Quantitative measurements of different types of blood cells, such as red blood cells, white blood cells, and platelets, which can indicate abnormalities or imbalances in the blood.

**4.4 Dataset Splitting and Model Selection:**

Dataset Splitting:

a. Split the dataset into training, validation, and testing sets.

b. The training set is used to train the machine learning model.

c. The validation set is used to tune the hyperparameters of the model.

d. The testing set is used to evaluate the performance of the final model

Model Selection:

Choose a suitable machine learning algorithm for classification, such as Random Forest, Support Vector Machines (SVM), or Neural Networks.

Perform feature selection to identify the most relevant K-mers for classification.

Train a separate model for each subset of K-mers to evaluate their individual performance.

Combine the models for the different subsets of K-mers to create an ensemble model.

Select an appropriate machine learning algorithm for prediction of blood cancer, such as confuse matrix, Random Forest, Support Vector Machines (SVM), or Neural Networks

**4.5 Interpretation of Results:**

Interpret the results of the model, including the identified features and their contribution to the prediction of blood cancer, Consider the clinical relevance of the identified features and the accuracy of the model in making prediction that whether it is a cancer or not cancer.

**5. OUTCOMES**

**5.1 Data Quality:** The reliability and representativeness of the data you gathered from NCBI and the accuracy of the bioinformatics analysis using BioPython are crucial. High-quality, well-curated data is essential for robust analysis.

Feature Selection: Identifying the most frequent k-mer associated with cancer is a significant step. However, the choice of features (in this case, k-mers) and how well they capture the underlying biology of cancer can impact the predictive power of your model.

**5.2 Mismatches:** Understanding the total number of mismatches is important, as it could provide insights into genetic variation. However, the interpretation of these mismatches requires careful consideration of their functional significance.

**5.3 Machine Learning Model:** The performance of your machine learning model will depend on the choice of algorithm, hyperparameter tuning, and the size and quality of your training dataset. It's essential to assess the model's accuracy, precision, recall, and other relevant metrics.

**5.4 Biological Context:** The identification of a k-mer associated with cancer is a significant finding, but the biological interpretation of this association is crucial. Understanding the functional relevance of the identified k-mer in the context of cancer biology is essential for drawing meaningful conclusions.

**5.5 Validation:** Ensure that your predictive model is validated on independent datasets to assess its generalizability. Overfitting to the training data can lead to poor performance on new, unseen data.

**5.6 Ethical and Clinical Implications:** Consider the ethical implications of your research, especially if it involves human data. Additionally, translating your findings into clinical applications requires careful consideration of the potential impact on patient care.

**6. TIMELINE OF THE PROJECT**

**6.1 Timeline Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Topics** | **Start date** | **End date** | **Days to complete** |
| Interaction | 24-Sep-23 | 24-Sep-23 | 1 |
| Review-0(finalize title) | 09-Oct-23 | 10-Oct-23 | 1 |
| Literature survey | 11-Oct-23 | 12-Oct-23 | 1 |
| Algo (SVM, k-NN) | 13-Oct-23 | 14-Oct-23 | 1 |
| Algo (neural networks, bayers, deep) | 14-Oct-23 | 16-Oct-23 | 2 |
| Objectives | 17-Oct-23 | 19-Oct-23 | 2 |
| Methodologies | 20-Oct-23 | 22-Oct-23 | 2 |
| Conclusion, reference | 23-Oct-23 | 24-Oct-23 | 1 |
| Making tables | 25-Oct-23 | 27-Oct-23 | 2 |
| Making ppt | 28-Oct-23 | 29-Oct-23 | 1 |
| Review-1 | 06-Nov-23 | 08-Nov-23 | 2 |
| Deciding program language(python) | 10-Nov-23 | 11-Nov-23 | 1 |
| Data collection (protein sequence) | 12-Nov-23 | 14-Nov-23 | 2 |
| Data collection | 15-Nov-23 | 17-Nov-23 | 2 |
| Making code for sequence | 19-Nov-23 | 21-Nov-23 | 2 |
| Making code for pattern | 22-Nov-23 | 24-Nov-23 | 2 |
| Executing both codes | 25-Nov-23 | 26-Nov-23 | 1 |
| Review-2 | 26-Nov-23 | 30-Nov-23 | 4 |

**6.2 Timeline Graph**

A blue background with a graph

Description automatically generated with medium confidence

**7. CONCLUSION**

We were able to identify and characterize the most common k-mers linked to blood cancer by using the BioPython tools in conjunction with a thorough examination of genomic data from the National Centre for Biotechnology Information (NCBI). In addition to identifying these important genomic regions, our analysis also measured the overall number of mismatches, shedding light on the genetic variants influencing the development of cancer.

Our capacity to anticipate blood cancer based on the discovered k-mers was substantially improved by the incorporation of machine learning methods. We have developed a prediction framework that shows promise for early identification and individualized treatment plans by utilizing the power of computer models. With this discovery, our understanding of the molecular causes of blood cancer will advance significantly, opening the door to more focused treatments and better patient outcomes.

The combination of bioinformatics and machine learning has shown to be a reliable method for deciphering the complexity of cancer genetics, as we commemorate the one-year mark of our research journey. The insights gleaned from this research not only advance the area of cancer research but may also have an influence on therapeutic procedures, ultimately pushing the boundaries of precision medicine.

**8. REFERENCES**

**8.1** Sathirapongsasuti et al. (2018) “Machine learning models for predicting prognosis in blood cancers”

**8.2** Kim et al. (2019) “Predicting risk of leukemia and lymphoma recurrence using machine learning”

**8.3** Esteva et al. (2017) “Early detection of blood cancers using deep learning on medical images”

**8.4** Zhang et al (2020) “Integration of multi-omics data for blood cancer prediction”

**8.5** Patel et al. (2021) “Predictive modelling of treatment response in blood cancer patients”

**8.6** Patel and Mishra “Automated Leukemia Detection Using Microscopic Images”

**8.7** Khalilabad and Hassanpour “[Employing image processing techniques for cancer detection using microarray images](https://www.sciencedirect.com/science/article/pii/S0010482516303328)”

**8.8** Rehman et al. “Classification of acute lymphoblastic leukemia using deep learning”

**8.9** Ebru Simsek, Hasan Badem & Ibrahim Taner Okumus “Leukemia Sub-Type Classification by Using Machine Learning Techniques on Gene Expression”

**8.10** Liuksiala & Thomas Edward “Blood Cancer Lineage Identification: A Machine Learning Approach”