ISSN 2723-6471

Optimizing Survival Prediction in Children Undergoing Hematopoietic Stem Cell Transplantation through Enhanced Chaotic Harris Hawk Deep Clustering

R. Arthi[1,*], G Maria Priscilla[2], Siti Sarah Maidin[3], Qingxue Yang[4]

[1]*Research Scholar and Assistant Professor Department of Computer Science, Sri Ramakrishna College of Arts and Science Coimbatore, India*

[2]*Associate Professor and Head Department of Computer Science, Sri Ramakrishna College of Arts and Science Coimbatore, India*

[3]*Faculty of Data Science and Information Technology (FDSIT), INTI International University, Nilai, Malaysia*

[4]*Faculty of Liberal Arts, Shinawatra University, Thailand*

(Received: August 25, 2024; Revised: October 6, 2024; Accepted: November 21, 2024; Available online: December 30, 2024)

Abstract

Cancer can impact individuals of all ages, including both children and adults. Diagnosing the pediatric cancer can be challenging due to its rarity. Typically, it is not recommended to screen for pediatric cancer as it may lead to potential harm to the children. One of the specialized treatments for pediatric cancer is Hematopoietic Stem Cell Transplant (HSCT). HSCT performs replacement of existing one's blood cells with the donor's bone marrow healthy cells. However, forecasting the survival rates following the pediatric HSCT is crucial and poses challenges in early detection. Many machine learning algorithms have been developed to predict the risk of transplant outcomes which depends on the type of disease or patient's comorbidity. In this work, the enhancement of survival prediction for children who have undergone hematopoietic stem cell transplantation (HSCT) is achieved through the introduction of a deep learning model that is based on behavioral characteristics. The primary aim of this model is to identify and differentiate between the patterns of malignancy, non-malignancy, and hematopoietic conditions within the dataset of bone marrow transplant patients. The existing unsupervised machine learning algorithms, performs clustering of instances with the randomly selected centroids, which often results in local optima and early convergence affects the accuracy rate. Hence, the present approach introduces Chaotic mapping Harris Hawk Optimization (CHHO) in order to enhance the conventional k-means clustering procedure due to its significantly reduced computational complexity. To understand the pattern of the bone marrow transplant dataset, the deep clustering model with its ability of auto encoder and decoder, discriminates the labelled instanced. With the inferred knowledge proposed CHHO with Deep clustering Model (CHHO-DCM) performs the effective clustering of instances with the advantage of both local and global optimization. The simulation outcomes have substantiated the effectiveness of the suggested CHHO-DCM model as it attains the highest level of precision when compared to the prevailing clustering models in predicting the survival of pediatric patients during Hematopoietic Stem Cell Transplantation (HSCT).s enduring HSCT.

Keywords: Hematopoietic Stem Cell Transplant, Chaotic mapping Harris Hawk Optimization, Deep Clustering Model

1. Introduction

Chemotherapy medications can damage the spongy bone marrow, responsible for producing blood cells. Due to their rapid growth, blood cells are particularly vulnerable to the negative impacts of these medications [1]. The majority of the stem cells in the bone marrow are killed by chemotherapy, but the cells eventually replenish [2]. A transplant of bone marrow is a surgical procedure in which healthy cells are used to replace your bone marrow and only option is using donor or body's replacement cells. A transplantation of bone marrow is often referred to as a hematopoietic cell transplant, or simply a stem cell transplant. Stem Cells hold immense significance in contemporary medicine due to their contributions towards the advancement of novel clinical methodologies. The investigation of stem cells can facilitate the comprehension of disease etiology and elucidate the potential of these cells in the realm of regenerative medicine. Leukemia, myeloma, lymphoma, along with other blood and autoimmune disorders that damage the bone

***Corresponding author: R. Arthi (arthi@irins.org)**
DOI: https://doi.org/10.47738/jads.v6i1.468

marrow can all be treated by transplantation [3]. Hematological stem-cell transplantation (HSCT) is the process of introducing multipurpose hematopoietic stem cells into a patient in order to allow them to proliferate inside of them and produce more normal blood cells [4]. These stem cells are often obtained from the bloodstream, bone marrow, or umbilical cord blood. Patients with life-threatening illnesses are the only ones eligible for HSCT because it is still a risky operation that has numerous potential consequences. The procedure's application has spread besides cancer to autoimmune illnesses as survival rates have grown. Despite being a life-threatening procedure, there is a chance that it could be fatal. HSCT is a specialized therapy for individuals suffering from particular tumors or other disorders which is not a surgical procedure. The goal of this type of treatment is to transplant healthy bone marrow into a patient after the patient's existing defective bone marrow has been treated with medication to eradicate the abnormal lymphocytes.

For certain patients with hematologic diseases, HSCT is a probably therapeutic operation. Even if the danger of transplantation has decreased recently, death and disability are still high, thus choosing who, what, and how often to do a transplant is crucial [5].

The need to employ algorithms based on machine learning (ML) for survival predictions is driven by the creation of vast and sophisticated archives that incorporate physiological and clinical information, as well as the demand for improved forecasting models. A branch of computational intelligence from the field of computing is called the field of ML. Instead of beginning with a preconceived approach, the underpinning paradigm allows the information build the model through discovering deeper trends. The data set includes both malignant and non-malignant pediatricians with various hematologic illnesses. Unaltered heterologous discrete recipient transplantation of hematopoietic stem cells was performed on every individual.

Though, there are many existing machine learning algorithms are used for survival prediction of pediatric patients. But the raw dataset comprised of missing values, inconsistent instances affect the accuracy rate of survival prediction of pediatric patients while undergoing hematopoietic stem cell transplantation. These problems are taken as the important factors which affects the accuracy of prediction models. Hence, in this proposed work, a novel Chaotic Harris Hawk optimization based deep clustering model is devised to overcome the vagueness in understanding the pattern of instances which exhibits malignant and non-malignant cases more precisely.

2. Related Work

Prior studies have consistently demonstrated a survival benefit associated with the utilization of younger donors compared to irrelevant donors. However, the influence of various factors has not been as apparent, and conventional statistical methods have occasionally failed to determine the effect of multiple donor factors on an individual based on their attributes.

Prompt and precise diagnosis plays a pivotal role in the successful management of diseases. Guncar et al [6] constructed two different machine learning algorithms to predict hematologic disease using the blood test parameters and admittance details of the patients respectively. The results showed that the model with reduced feature set significant produce better accuracy compared with the other model. But the reliability of testing the unknown patterns are not discussed in this work.

The diagnosis of malignancy in hematological management is survey in detail by Radakovichet al [7] using the machine learning models. The review report underscores the significance of machine learning, which constructs a strategic plan for the fields of genomics, pathology, and the analysis of healthcare datasets pertaining to patients who are suspected to be in the malignant phase. They also discussed about the significant of machine learning that creates awareness to physicians.

Choiet al [8] developed a novel prediction model using machine learning algorithms for prolonged survival after HCT in patients with the malignancies of hematologic. The conditioning regimens-based donor selection for transplantation is also discussed in their work.

After conducting hematopoietic stem cell transplantation, the risk of mortality rate is very high, to determine the predict the survivability rate, Aziz Nazha1et al., [9] introduced a random survival forest method. This work explores the importance of identifying whether the transplantation offers benefit or not to the patients before undergoing HCT.

Ying Li et al [10] in their work used three various machine learning algorithms namely, support vector machine, logistic regressions and boosting decision tree for predicting the availability of the donors. The results show that the boosted classifier achieves better accuracy.

Long XiangEt al., [11] devised an artificial intelligence based early warning algorithm to identify the malignancies after hematological transplant. The XGBoost algorithm is deployed for prediction of septic shock at its early stage by determining pediatric sequential organ failure score.

El Alaoui et al., [12] conducted a detailed analysis and review in the area of haematological management by explored the importance of deep learning and machine learning applications in blood cancer research. This work focuses on discovering the patient's cancer severity to improvise further research on therapies related to blood cancer.

From the above works discussed in this section, most of the algorithms are involved in survivability prediction of patient's after gone haematological transplant. But, still the issue of presence of vague and inconsistent pattern in dataset is not analysed and focused in the previous works which affects the accuracy rate of survivability prediction children enduring Hematopoietic Stem Cell Transplantation. Hence, in this work to discriminate the malignancy, non-malignant and hematologic instances are accomplished by introducing the behavioural based metaheuristic algorithm with deep learning algorithm, it is discussed in the following sections.

3. Methodology

In this phase, to conduct effective survival prediction of Children Enduring Hematopoietic StemCell Transplantation the quality of the dataset is enhanced. The collected raw dataset comprised of missing values, so before performing pattern discovery, in this work fuzzy K-Nearest Neighbour is used for imputing missing values [13]. With the complete dataset after imputation is used for determining the patterns of instances belongs to Non-malignant or Malignant or Hematologic instances by devising a novel algorithm known as Chaotic Harris Hawk Optimization based deep clustering model (CHHO-DCM) [14]. The proposed model CHHO-DCM, with the knowledge of Harris hawk food searching behaviour centroids are selected and the similar instance are clustered with the deep clustering model. Thus, the pattern of normal children and children with hematologic symptoms are discriminated in this phase. Figure 1 show the Architecture of Enriched Chaotic Harris Hawk Optimized Deep Clustering for survival prediction.

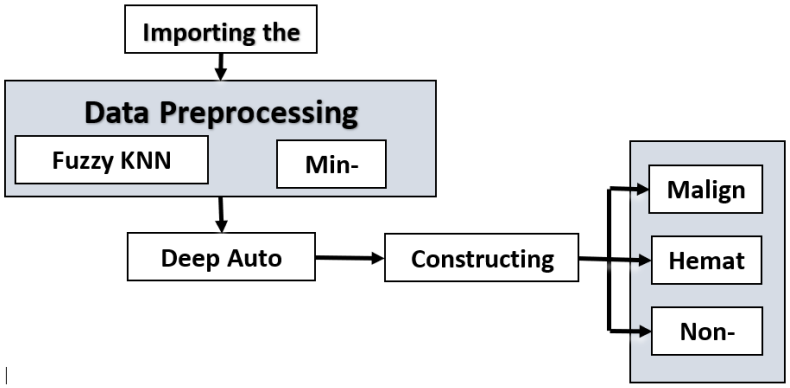


Figure 1. Architecture of Enriched Chaotic Harris Hawk Optimized Deep Clustering for survival prediction

3.1. Missing Value Imputation using Fuzzy KNN

The Fuzzy K Nearest Neighbour (FKNN) receives two sets of separated instances and produces an additional pair of interpolated cases. The FKNN method examines the collection of entire cases and provides the K perfect examples that are closest to it [15]. The present unfinished instance should be comparable. FKNN generates a fuzzy set with m values and an association score of 1 for every value that is missing in an incomplete instance, while m is the total number of total occurrences [16]. FKNN, in simple terms, considers that the values that are not present may correspond to any combination of earlier encountered values incomplete instances. Following that, FKNN scans the entire set of instances once and determines the aggregate proximity among the present partial instance and the instances with complete data.

 ISSN 2723-6471

To modify the disparity among the present, whole, and partial instances, FKNN changes the set of fuzzy data by one component on each iteration. Following the completion of the dissimilarity computation, FKNN classifies every component in the fuzzy set in descending order depending on the membership value. Finally, FKNN chooses the top nearest instance's value to fill the value that is missing in an unfilled instance. The FKNN method is repeated for every case in the partial cases set to generate a fresh imputation subset.

3.2. Data Pre-Processing

The Bone marrow Transplant dataset is pre-processed by applying min-max normalization by converting the range of dataset values to fall under same range i.e. zero and one. The formula for normalizing is as follows

$$\text{Normz}(y) = \frac{y-\min(Y)}{\max(Y)-\text{Min}(Y)} \text{----Equation} \quad (1)$$

Where y mentions to a single attribute value and Y denotes to the complete range of values of a specific attribute's overall instances range with minimum and maximum values.

3.3. Chaotic Harris Hawk Optimized Deep Autoencoder based Clustering Algorithm

This proposed work has two key processes for detecting a recurring trend among instances of training set and test dataset. The training set is used to train a deep autoencoder using an encoding system and a decoder in the training stage. A compressed input vector is sent into a multilayer deep encoder with a low dimensional learnt representation in this case. This acquired model is then given into a decoder, which attempts to return an output with the identical dimension as the input data [17]. This autoencoder's training procedure tries to recreate the data provided as much as feasible. The autoencoder is applied to the dataset in the next clustering stage. The encoder output is then input into a conventional K-Means algorithm for grouping. The chaotic Harris Hawk optimization for selection of centroids [18]. The learned low dimensional representation vector contains key information of the given input, and thus yield better clustering results. Figure 2 shows an overview of our deep autoencoder-based clustering framework.

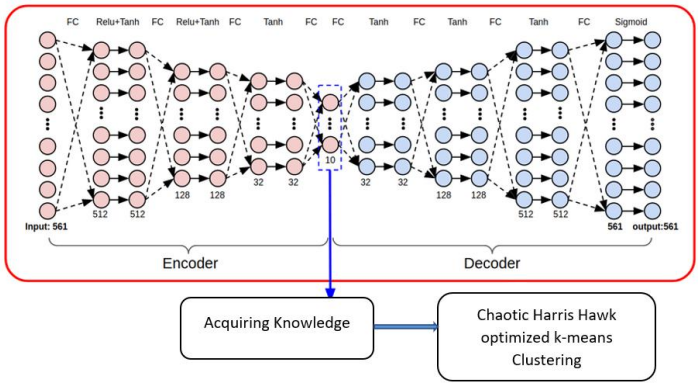


Figure 2. Overview of our deep auto encoder-based clustering framework.

Deep Cluster, a clustering method that jointly learns the parameters of a neural network and the cluster assignments of the resulting features. Deep Cluster iteratively groups the features with a standard clustering algorithm, k-means, and uses the subsequent assignments as supervision to update the weights of the network model is not as complex as some of the advanced neuron networks. The reason is that we do not want our model to over-fit in two-folds. First, to avoid the proposed model to over-fit on the training dataset over the testing dataset. Second, we do not want our model to over-fit on the reconstruction problem it-self over the clustering problem. Thus, we select a model of reasonable median complexity.

The encoder aims to encode or compress the input data into a smaller size representation, and at the same time preserve as much key information as possible. As shown in figure 1, the encoder consists of 8 layers, include the input layer and the learned representation output layer. Here the input layer is being normalized such that all its values is in the range of (0, 1). Specifically, from the beginning, each larger layer is fully connected to the next smaller layer followed by a couple of activation layers.

 ISSN 2723-6471

There are mainly two types of activation layers, Relu and Tanh, as shown in Equation 1 and 2. Adding the Relu layers could introduce non-linearity to our model, making it more robust against non-linear input data. The Tanh layer, on the other hand, could transform the data into a normalized range of (−1, 1), to alleviate the gradient vanishing/exploding problem

$$\text{RLU (y)} = \text{MAX}(0,y) \tag{2}$$

$$Cosh(y) = \frac{e^{y}+e^{-y}}{2} \tag{3}$$

$$Sinh(y) = \frac{e^{y}-e^{-y}}{2} \tag{4}$$

$$tanh(y) = \frac{Sinh\,(y)}{Cosh(y)} = \frac{e^{y}+e^{-y}}{e^{y}-e^{-y}} \tag{5}$$

The decoder aims to decode or decompress the encoded output to reconstruct the original input data as much as possible. It contains nine layers, including the input layer, which is the output of the encoder, and the final output layer. Specifically, each smaller layer is fully connected to the next larger layer followed by a Tanh activation layer. In addition, the decoder has a Sigmoid activation layer (shown in Equation 3) at the final stage to enforce the output values lie into the range of (0, 1)

$$\text{Sig(y)} = \frac{1}{1+e^{-y}} \tag{6}$$

Clustering-weighted MSE Loss while the goal of the classic autoencoder is to reconstruct the original input as much as possible, it counts each input feature value equally. However, it is possible that each individual input feature contributes differently to the final clustering results.

The k-means clustering performance is boosted by adopting the Harris Hawk Optimization algorithm for selection the centroids by balancing both local and global optima to avoid the early convergence.

3.4. Chaotic Harris Hawks optimizer Method

Harris Hawks Optimization algorithm is a social approach that is impacted by community perception, and its main advantages are collegiality and unexpected attack attacking brilliance. As depicted in figure 3, during the course of searching for the subject of the hunt, every predator or hawk pursues the prey from multiple vantage points. This strategy is adopted to locate the optimal target in an uncertain domain of search of centroid selection in deep clustering.

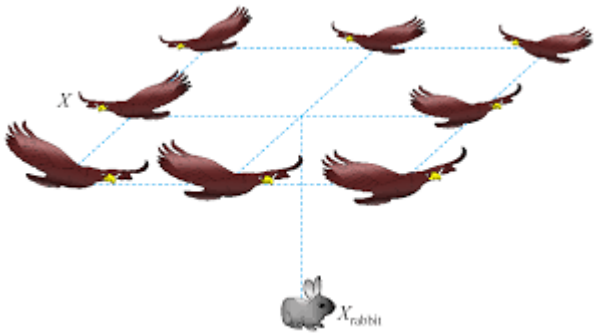


Figure 3. Harris Hawk Prey searching

Harris hawk hunting methods prompted the development of the HHO algorithm. Harris hawks are considered one of the most intelligent animals on the planet, and they often hunt in groups. As soon as they capture prey, they distribute it to the entire herd of hawks. When creating HHO, the prey is supposed to be a rabbit. Harris hawks spend a long time searching in the dunes for a rabbit. When a prey is found at the terminus of this hunt, the herd proceeds to follow it and attack when the time comes. Hawks plan and carry out their attacks in tandem. This form of hunting necessitates a very astute plan. The early attacks had been intended to weaken the rabbit's resistance [19]. The hawks then launch their primary attacks in order to capture the encircled rabbit with little defence resistance. Because the rabbit rushes out of its usual place of refuge and looks for different areas of concealment while the time it is attacked, it seems to be simpler to trap.

 ISSN 2723-6471

Finally, the hawks effortlessly catch the exhausted rabbit without expending much energy. This tactic demonstrates that Harris hawks have fine-tuned their hunting methods. The HHO algorithm is divided into two stages based on its discovery and attack techniques.

Exploration phase: In the course of their exploration, Harris hawks roost in an area to seek for their prey, which in this case is a rabbit. Each hawk symbolises a potential solution, while the bunny represents the best conceivable answer. The hawks get nearer to the prey at each level, which is the ideal solution. Harris hawks have two roosting techniques. In the model, these two techniques are provided with the same probability (r). The first technique is to perch based on the prey's location ($r < 0.5$). The second is feeding at close quarters ($r \geq 0.$). The 2ndapproach tries to streamline flock members' interactions while foraging. Formula denotes these two approaches are shown in equation

$$\mathrm{Z(p+1)} = \{\mathrm{Z_{rnd}(p)} - \gamma_1|\mathrm{Z_{rnd}(p)} - 2\gamma_2\mathrm{Z(p)}|\mathrm{r} \geq 0.5\mathrm{Z_{PRY}(p)} - \mathrm{Z_v(p)}) - \gamma_3(\mathrm{LB} + \gamma_4(\mathrm{UB} - \mathrm{LB}))\mathrm{r}0.5 \quad (6)$$

The position vector of the hawk (search agent) is $Z(p+1)$ at the period p+1, the next generation of rabbit is Y_{rbt}. The recent position of search agent is Z(p), random numbers are r and $\gamma_1, \gamma_2, \gamma_3. \gamma_4$whose values lies between (0,1). These values are updated during each iteration, the lower and the support bound for the search agent is defined by the parameters UB and LB. $Z_{rnd}(p)$is the random search agent chosen and the current crowd's mean position is signified as Z_v.The average position of the searching agent is formulated as \

$$\mathrm{z_v(t)} = \frac{1}{\mathrm{N}}\sum_{i=1}^{N}\mathrm{Z_i(p)} \quad (7)$$

Where $Z_i(p)$ is the present location of each searching hawk, s is the iteration and N is the total number of search agent population.

Exploitation phase: During this phase the prey's energy is determined using the formula

$$\mathrm{Eg} = 2\mathrm{Eg_0}(1 - \frac{\mathrm{t}}{\mathrm{Tr}})\text{---} \quad (8)$$

Where Eg, Tr and Eg0 are the target; seschaping energy, total generations and energy's initial state as shown in the figure 4.

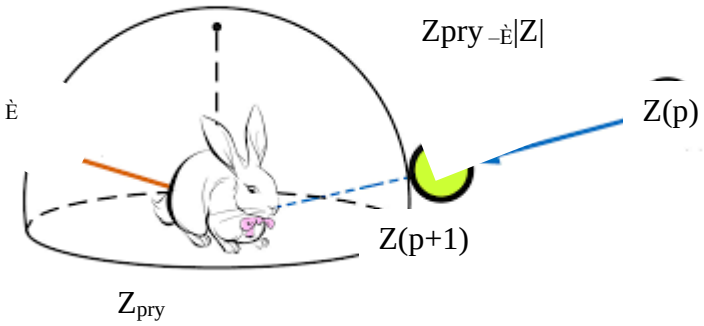


Figure 4. Position of prey with computation of energy

Exploitation phase: There are two behaviors that must be simulated during the exploitation phase [20]. In soft besiege is the first behaviour, in which the rabbit's energy is still strong and it can flee quickly. In this case, Harris hawks try to follow it slowly and keep an eye on it until it becomes fatigued. While in hard besiege, the prey in this activity is exhausted and unable to flee. As a result, in this mode, the Harris hawks form closed circles to launch a surprise attack.

While $Eg \geq 0.5\ and\ er \geq 0.5$, this condition signifies that the prey or rabbit has relatively high escaping energy (Eg) and the change of successful escape (er) is greater than 50%. Then Harris hawk exhibits soft besiege behaviour, the search agent's behaviours are accomplished using the following formula

$$\mathrm{Z(p+1)} = \Delta\mathrm{Z(s)} - \mathrm{Eg}|\mathrm{JZ_{pry}(p)} - \mathrm{Z(p)} \quad (8)$$

$$\Delta\mathrm{Z(p)} = \mathrm{Z_{pry}(p)} - \mathrm{Z(p)} \quad (9)$$

 ISSN 2723-6471

The random jump strength of rabbit based on its escaping procedure is represented using J = 2(1-γ_5), $\Delta Y(p)$refers to variation among location vector of the rabbit, γ_5 is a random number and s is the current position in the generation. To simulate the rabbit's nature, the value of J will be updated randomly during each iteration.

When $Eg \geq 0.5\ and\ er < 0.5$ It indicates that the rabbit has high energy. However, the likelihood of successfully fleeing is low, Harris hawk performs progressive soft besiege with rapid dives. The next move of the hawks is mathematically updated as

$$B = Z_{pry}(p) - \dot{E}|Z_{pry}(p) - Z(p)| \quad (10)$$

In Hard besiege the current position of all agents are updated as

$$Z(p+1) = Z_{pry}(p) - Eg|\Delta Z(p)| \quad (11)$$

The hawks will then determine which dive is better by comparing the present position to the prior dive. The hawks will take the previous dive if it is better. Otherwise, the hawks will use the levy flight to make a new dive.

$$H = B + R * \beta(M) \quad (12)$$

B is the present location, R is the random vector, M is the dimension of the problem and β is the levy flight. In both soft besiege and hard besiege behaviour, it is necessary to replace the positions of all members.

$$Z(p+1) = \{Z, \beta(X) < F(Y(s)H, \beta(H) < F(Y(s)\text{---} \quad (13)$$

When dealing with an optimization problem, it is frequently necessary to create a function with an objective that can model the problem's needs. The target function is usually written in such a way that it converts the issue of optimization into a matter of minimization. Using the Gradient Descent Optimizer (GDO) to attain an overall minimum in minimization issues. However, this is true for concavity issues. GDO may be trapped by the local minimum in cases with multiples of the local minimum.

4. Experimental Results

This section discusses the performance analysis of the newly developed algorithm Chaotic Harris Hawk Optimized Deep Clustering Model (CHHO-DCM), their parameter is shown in the figure 4. The simulation results are obtained using python software. The CCHO-DCM for survival prediction of Children Enduring HSCT using Bone marrow Transplant dataset is compared with three different unsupervised algorithms namely k-Means, Fuzzy C Means and traditional deep clustering model. Table 1 show the parameters used in this study.

Table 1. Parameters of Deep Encoder and Decoder

Layer (type)	Output Shape	Param #
input (InputLayer)	(None, 50)	0
encoder_0 (Dense)	(None, 500)	25,500
encoder_1 (Dense)	(None, 500)	250,500
encoder_2 (Dense)	(None, 2000)	1,002,000
encoder_3 (Dense)	(None, 10)	20,010

4.1. Evaluation Metrics Used

$$\text{Precision (Model)} = \frac{\text{Sucessfully clustered instances}}{\text{Total instances of positive and negative observations}} \quad (14)$$

$$\text{Recall (Model)} = \frac{\text{Sucessfully clustered instances}}{\text{Missed Instances}}\text{---} \quad (15)$$

Table 2 presents the performance comparison of four clustering approaches: K-Means, Fuzzy C-Means (FCM), Deep Clustering, and CHHO-DCM. The results highlight the percentage of correctly and incorrectly clustered instances for each model. K-Means achieves a clustering accuracy of 77.1%, with 22.9% of instances misclassified, indicating relatively lower performance. FCM improves upon K-Means with 81.67% correctly clustered instances and a reduced error rate of 18.33%. Deep Clustering further enhances accuracy, achieving 87.53% correct clustering and a

 ISSN 2723-6471

significantly lower error rate of 12.47%. Among all models, CHHO-DCM (Clustered Hybrid Heuristic Optimization with Deep Clustering Model) outperforms the others, achieving an impressive 95.82% accuracy and a minimal error rate of 4.18%. These results demonstrate the superior performance of CHHO-DCM, showcasing its capability to deliver highly accurate clustering results by leveraging a hybrid heuristic and deep

Table 2. Results of four Clustering Models

Clustering Approaches	Correctly Clustered Instances	Incorrectly Clustered Instances
K-Means	77.1	22.9
FCM	81.67	18.33
Deep Clustering	87.53	12.47
CHHO-DCM	95.82	4.18

Figure 5 and figure 6 displays the results produced by applying the full feature set with deep clustering using Harris hawk optimization produced the highest rate of correctly clustered rate compared with the existing deep clustering, Fuzzy C Means Clustering, Density based Clustering. For centroid selection the existing models performs random selection initially, and based on the Euclidean distance and the highest number of neighbouring instances, cluster heads are reassigned in the consecutive clustering process. In the proposed model, the selection of initial centroids is done with the help of chaotic Harris hawk optimization, where the initial selection optimizes the clustering process. With the acquired knowledge by applying Deep autoencoder model, the proposed CHHO-DCM produced better result compared with the existing clustering models.

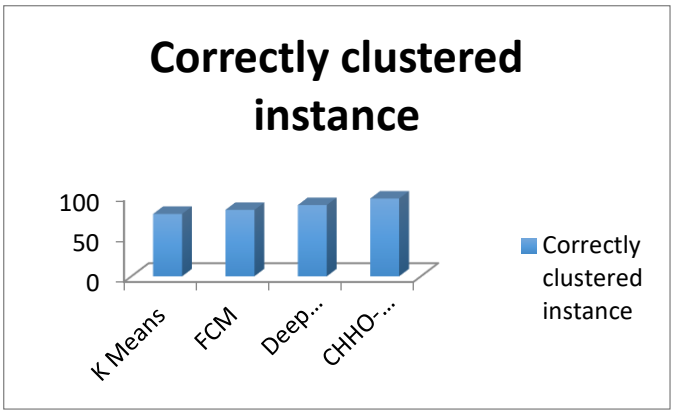


Figure 5. Performance Comparison based on correctly clustered

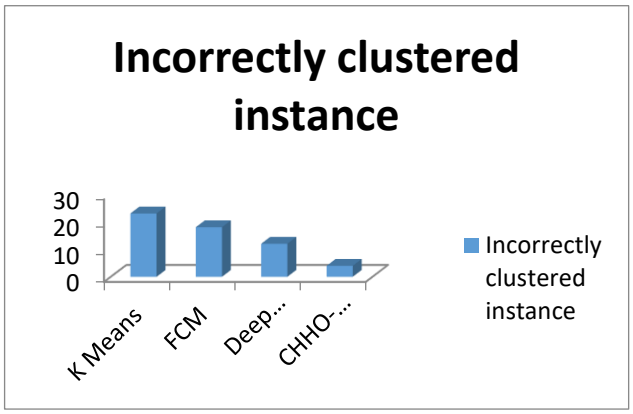
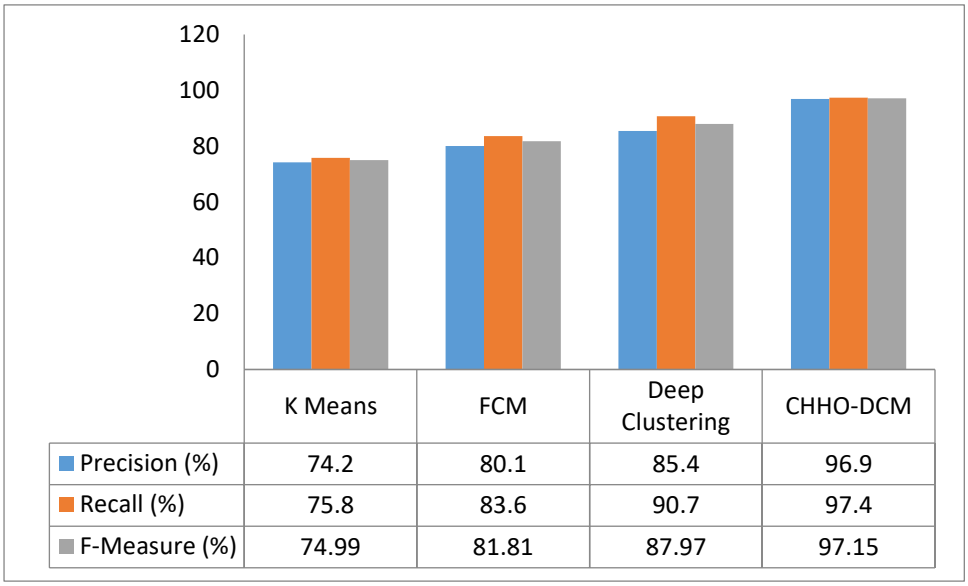


Figure 6. Performance comparison based on incorrectly clustered

Figure 7 explores precision rate of the clustering models to discover the three different patterns namely benign, malignant and Hematologic Instances in predicting survival of Children Enduring Hematopoietic Stem Cell Transplantation. The proposed CHHO-DCM achieves higher rate of accuracy compared to the existing clustering models like k-means, fuzzy c Means and Conventional Deep Clustering. In CHHO-DCM, the problem of local optimization in selection of centroids in the presence of vague instances, is handled by the chaotic Harris Hawk optimization-based clustering improves the clustering rate of similar pattern inhibiting instances more precisely while compared with conventional deep clustering, Fuzzy C Means and k-means clustering for categorizing the pattern of children's with hematologic symptoms and healthy children.

The recall rate of the CHHO-DCM is higher than the other three clustering models because of handling the missing values using the uncertainty theory of fuzzy concept to determine the closest complete instances and imputing the missing instances. The improvised complete dataset is further processed by the proposed clustering model to gain the highest rate of recall compared with the other existing clustering models.

 ISSN 2723-6471



	K Means	FCM	Deep Clustering	CHHO-DCM
Precision (%)	74.2	80.1	85.4	96.9
Recall (%)	75.8	83.6	90.7	97.4
F-Measure (%)	74.99	81.81	87.97	97.15

Figure 7. Performance based on precision rate, Recall and F-measure

The rate of F-Measure is obtained with the influence of precision and recall values, so the proposed CHHO-DCM achieves highest value of 97.15% while the other clustering models obtains the lower rates. The optimized centroid selection and balancing the local and global searching chaotic mapping help the proposed CCHO-DCM to achieve best highest accuracy compared with other state of arts in determining the patterns of instances such as Non-malignant, Malignant and Hematologic instances.

5. Conclusion

This study emphasizes the significance of predicting the survival outcomes of children undergoing Hematopoietic Stem Cell Transplantation (HSCT) and aims to differentiate individuals into categories of malignancy, non-malignancy, and hematologic instances. Existing clustering models struggle with handling inconsistent instances that are challenging to predict within specific clusters. Addressing this issue, we introduce the Chaotic Harris Hawk Optimized Deep Clustering Model (CHHO-DCM). This model utilizes a training dataset to thoroughly understand instance patterns through the application of deep autoencoder and decoder classifiers. Subsequently, the testing dataset instances are clustered into relevant groups using Chaotic Harris Hawk Optimized k-means clustering. The primary contribution lies in addressing local and global imbalances in seeking centroids and clustering instances for predicting the survivability of children undergoing HSCT. The results demonstrate that CHHO-DCM outperforms existing clustering models in terms of accuracy. Future efforts will focus on identifying features influencing the survivability prediction of children with HSCT to enhance the model's reliability.

6. Declarations

6.1. Author Contributions

Conceptualization: R.A., G.M.P., S.S.M., Q.Y.; Methodology: S.S.M.; Software: R.A.; Validation: R.A., S.S.M., and Q.Y.; Formal Analysis: R.A., S.S.M., and Q.Y.; Investigation: R.A.; Resources: S.S.M.; Data Curation: S.S.M.; Writing Original Draft Preparation: R.A., S.S.M., and Q.Y.; Writing Review and Editing: S.S.M., R.A., and Q.Y.; Visualization: R.A. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. K. K. Sathish, M. Chaturvedi, P. Das, S. Stephen, and P. Mathur, "Cancer incidence estimates for 2022 and projection for 2025: Result from National Cancer Registry Programme, India," *Indian Journal of Medical Research*, vol. 156, no. 4and5, pp. 598–607, May 2022.

[2] V. Acharya, V. Ravi, T. D. Pham, and C. Chakraborty, "Peripheral Blood Smear Analysis Using Automated Computer-Aided Diagnosis System to Identify Acute Myeloid Leukemia," *IEEE Transactions on Engineering Management*, vol. 2021, no. 8, pp. 1–14, 2021.

[3] D. T. Chu, T. T. Nguyen, N. L. B. Tien, D. K. Tran, J. H. Jeong, P. G. Anh, V. V. Thanh, D. T. Truong, and T. C. Dinh, "Recent Progress of Stem Cell Therapy in Cancer Treatment: Molecular Mechanisms and Potential Applications," *Cells*, vol. 9, no. 3, pp. 1–15, Mar. 2020.

[4] S. O. Ciurea, P. Kongtim, and O. Hasan, "Validation of a Hematopoietic Cell Transplant-Composite Risk (HCT-CR) Model for Post-Transplant Survival Prediction in Patients with Hematologic Malignancies," *Clinical Cancer Research*, vol. 26, no. 10, pp. 2404–2410, May 2020.

[5] M. Beksac, S. Iacobelli, and L. Koster, "An early post-transplant relapse prediction score in multiple myeloma: A large cohort study from the chronic malignancies working party of EBMT," *Bone Marrow Transplantation*, vol. 58, no. 6, pp. 916–923, Jun. 2023.

[6] G. Guncar, M. Kukar, M. Notar, M. Brvar, P. Cernelc, M. Notar, and M. Nota, "An application of machine learning to haematological diagnosis," *Scientific Reports*, vol. 8, no. 411, pp. 1–7, Jan. 2018.

[7] A. Kumar, R. S. Umurzoqovich, N. D. Duong, P. Kanani, A. Kuppusamy, and M. Praneesh, "An intrusion identification and prevention for cloud computing: From the perspective of deep learning," *Optik*, vol. 270, no. 11, pp. 1–12, Nov. 2022.

[8] N. Radakovich, M. Nagy, and A. Nazha, "Machine learning in haematological malignancies," *The Lancet Haematology*, vol. 7, no. 7, pp. e541–e550, Jul. 2020.

[9] E. J. Choi, T. J. Jun, H. S. Park, J. H. Lee, K. H. Lee, Y. H. Kim, Y. S. Lee, Y. A. Kang, M. Jeon, H. Kang, J. Woo, and J. H. Lee, "Predicting long-term survival after allogeneic hematopoietic cell transplantation in patients with hematologic malignancies: Machine learning-based model development and validation," *JMIR Medical Informatics*, vol. 10, no. 3, pp. 1–20, Mar. 2022.

[10] A. Nazha, Z. H. Hu, T. Wang, and R. C. Lindsle, "A personalized prediction model for outcomes after allogeneic hematopoietic cell transplant in patients with myelodysplastic syndromes," *Biology of Blood and Marrow Transplantation*, vol. 26, no. 11, pp. 2098–2104, Nov. 2020.

[11] Y. Li, A. Masiliune, D. Winstone, L. Gasieniec, P. Wong, H. Lin, R. Pawson, G. Parkes, and A. Hadley, "Predicting the availability of hematopoietic stem cell donors using machine learning," *Biology of Blood and Marrow Transplantation*, vol. 26, no. 11, pp. 1406–1413, Nov. 2020.

[12] L. Xiang, H. Wang, S. Fan, W. Zhang, H. Lu, B. Dong, S. Liu, Y. Chen, Y. Wang, L. Zhao, and L. F. Liebin, "Machine learning for early warning of septic shock in children with hematological malignancies accompanied by fever or neutropenia: A single center retrospective study," *Frontiers in Oncology*, vol. 11, no. 678743, pp. 1–12, Jun. 2021.

[13] B. M. Bai, N. Mangathayaru, and B. Rani, "Modified K-Nearest Neighbour Using Proposed Similarity Fuzzy Measure for Missing Data Imputation on Medical Datasets (MKNNMBI)," *International Journal of Fuzzy Systems Applications*, vol. 11, no. 1, pp. 1–15, 2022.

[14] Z. Elgamal, N. Yasin, M. Tubishat, M. Alswaitti, and S. Mirjalili, "An Improved Harris Hawks Optimization Algorithm With Simulated Annealing for Feature Selection in the Medical Field," *IEEE Access*, vol. 8, no. 1, pp. 186638–186652, 2020.

[15] Z. Bian, C. Vong, P. Wong, and S. Wang, "Fuzzy KNN method with adaptive nearest neighbors," *IEEE Transactions on Cybernetics*, vol. 52, no. 1, pp. 5380–5393, 2020.

[16] M. M. Kumbure, P. Luukka, and M. Collan, "A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean," *Pattern Recognition Letters*, vol. 140, no. 1, pp. 172–178, 2020.

[17] Y. Opochinsky, S. E. Chazan, S. Gannot, and J. Goldberger, "K-Autoencoders Deep Clustering," *in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2020, no. 1, pp. 4037–4041, 2020.

[18] T. Singh, "A chaotic sequence-guided Harris hawks optimizer for data clustering," *Neural Computing and Applications*, vol. 2020, no. 1, pp. 1–15, 2020.

[19] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. M. Mafarja, and H. Chen, "Harris hawks optimization: Algorithm and applications," *Future Generation Computer Systems*, vol. 97, no. 1, pp. 849–872, 2019.

[20] M. Al-Betar, M. Awadallah, A. Heidari, H. Chen, H. Al-Khraisat, and C. Li, "Survival exploration strategies for Harris Hawks Optimizer," *Expert Systems with Applications*, vol. 168, no. 1, pp. 1-12, 2020.