Proceedings of the Federated Conference on Computer Science and Information Systems pp. 15–20
DOI: 10.15439/2018F52
ISSN 2300-5963 ACSIS, Vol. 15

Kestrel-based Search Algorithm (KSA) for parameter tuning unto Long Short Term Memory (LSTM) Network for feature selection in classification of high-dimensional bioinformatics datasets.

Israel Edem Agbehadji
ICT and Society Research Group Department of Information Technology Durban University of Technology, Durban, South Africa. Email: *21648757@dut4life.ac.za*

Richard Millham
ICT and Society Research Group Department of Information Technology Durban University of Technology, Durban, South Africa. Email: *richardm1@dut.ac.za*

Simon James Fong
ICT and Society Research Group Department of Computer and Information Science University of Macau, Macau, SAR Email: *ccfong@umac.mo*

Hongji Yang
Department of Computer Science University of Leicester Leicester, UK Email: hongji.yang@gmail.com

***Abstract*—Although deep learning methods have been applied to the selection of features in the classification problem, current methods of learning parameters to be used in the classification approach can vary in terms of accuracy at each time interval, resulting in potentially inaccurate classification. To address this challenge, this study proposes an approach to learning these parameters by using two different aspects of Kestrel bird behavior to adjust the learning rate until the optimal value of the parameter is found: random encircling from a hovering position and learning through imitation from the well-adapted behaviour of other Kestrels. Additionally, deep learning method (that is, recurrent neural network with long short term memory network) was applied to select features and the accuracy of classification. A benchmark dataset (with continuous data attributes) was chosen to test the proposed search algorithm. The results showed that KSA is comparable to BAT, ACO and PSO as the test statistics (that is, Wilcoxon signed rank test) show no statistically significant differences between the mean of classification accuracy at level of significance of 0.05. However, KSA, when compared with WSA-MP, shows a statistically significant difference between the mean of classification accuracy.**

***Index Terms*—kestrel-based search algorithm, deep learning, random encircling, long short term memory network.**

I. Desire Models

THE CONCEPT of big data may be characterized by volume, velocity, value, veracity and variety. The volume relates to the amount of data that has to be processed within a given time; velocity relates to how fast incoming data need to be processed and how quickly the receiver of information needs the results from the processing system [1]; and the value is what a user will gain in terms of insight from the data analysis; the variety is the different structures that data may take such as text and images while the veracity is authenticity of the data source. In order to manage effectively these aspects of big data, an important step is to reduce the volume of dataset by selecting relevant features for classification. However, this may not be achieved without tuning different parameters that fit the data to select relevant features and ensure accurate classification. This paper proposes a search strategy for classification that is based on the behaviour of kestrel bird (to discover the optimal weight parameter) and deep learning network (for classification of features).

The related work is presented in Section II. Section III describes the behaviour of the Kestrel bird with its mathematical modeling and algorithm. Section IV outlines the experimental setup; and provides experimental results with comparative meta-heuristic algorithms. Conclusions and future work are given in Section V.

II. Related Work

Feature selection is the process of selecting relevant features from large number of features in a dataset, while ignoring the rest of features that have little value on the output feature set. The feature selection methods are categorized into the filter method (that is classifier-independent) [2], wrapper method (that is classifier-dependent) [2] and embedded method [3]. However, when big data is involved, it results in high computational cost in training and selection of features [4]. This challenge led to the concept of deep learning which historically originated from artificial neural network [5].

A. Deep learning network

Deep learning is a sub-field of machine learning that is based on learning several levels of representations, corresponding to a hierarchy of features where higher-level features are defined from lower-level ones, and the same lower level features can help to define many higher-level features [5, 6]. It has been indicated in [7] that deep learning is a statistical technique that help in classifying patterns based on sampled data using neural networks with multiple layers. In principle, deep learning uses multiple hidden layers of non-linear processing that is hierarchical; and uses different parameters to learn from hidden layers using algorithms (such as back-propagation algorithms) with large amounts of available training data [8]. Deep learning methods for classification are deep discriminative models/supervised-learning (e.g., deep neural networks (DNN), recurrent neural networks (RNN), etc.) and generative/unsupervised models (e.g., deep belief networks (DBN), etc.). Deep neural network (DNN) sometimes referred at as DBN is a multilayer network with many hidden layers, whose weights are fully connected and initialized (pre-trained) using stacked RBMs or DBN [9]. Recurrent neural networks (RNN) is a discriminative model but also has been used as generative model where output results from a model represents the predicted input data. When RNN is used as a discriminative model, the output results from the model is a label sequence associated with input data sequence [9]. The learning of parame-

ters in RNN are improved by information flow in bi-directional RNN and by a cell with LSTM (long short-term memory where cells are responsible for remembering parameters within a time interval) [6] which are the building units for layers of RNN. The RNN composed of LSTM units is often referred to as LSTM network. However, the challenge with the RNN is that when training neural network for deep learning classification problems, the back-propagated gradients approach that is often used either grows or shrinks at each time step, so over many time steps it typically explodes or vanishes [10]. Building a classification model from deep learning techniques integrated with metaheuristic search methods (also referred to as random search strategy as earlier mentioned) enhances accuracy/quality to select useful and relevant features [11] in a dataset. The advantage of meta-heuristic search method is the use of random search strategy to avoid being trapped in local optima when the search space grows exponentially.

B. *Meta-heuristic algorithms*

Among the random search/meta-heuristic algorithms for feature selection in classification problems are Genetic algorithm (GA) [12], Ant Colony Optimization (ACO) [13], Particle Swarm Optimization (PSO) [14], BAT [15] and Wolf Search Algorithm (WSA) [16].

Genetic algorithms is an evolutionary approach that is based on survival of the fittest. Genetic algorithm has the biological principle that species live in a competitive environment and their continuous survival depends on the mechanics of "natural selection" (Darwin, 1868 as cited by [12]) in which an element or chromosomes in the genetic structure is represented by a binary string. A genetic algorithm is an adaptive search procedure which involves the use of operators such as crossover, mutation and selection methods to find a global optimal results/solution by optimizing an objective function/fitness function.

The Ant Colony Optimization (ACO) [13] is a meta-heuristics search method that is inspired by the foraging behavior of real ants in their search for the shortest paths to food sources. When a source of food is found, ants deposit pheromone to mark their path for other ants to traverse. Pheromone is an odorous substance that is used as a medium for indirect communication among ants. The quantity of pheromone depends on the distance, quantity and quality of food source. However, pheromone substance tends to decay or evaporate with time. While a lost ant that moves at random detects a laid pheromone, it is likely that it will follow the path to reinforce the pheromone trails by further depositing some amount of the trail substances while this path leads to a desired outcome. If the path does not lead to a desired outcome, it is no longer followed and the pheromone evaporates in time until it is no longer detectable. Thus, ants make probabilistic decisions on updating their pheromone trail and local heuristic information in order to explore larger search areas. The ACO has been applied to solve many optimization related problems, including data mining, where it was shown to be efficient in finding best possible solutions. ACO, when applied to feature selection, improves on performance of feature selection by finding the best possible path.

The Wolf Search Algorithm (WSA) [16], is bio-inspired heuristic optimization algorithm which is based on wolf preying behavior. The behaviour of wolves includes the ability to hunt independently by remembering their own trait (meaning wolves have memory); ability to only merge with its peer when the peer is in a better position (meaning there is trust among wolves to never prey on each other); ability to escape randomly upon appearance of a hunter; and the use of scent marks as a way of demarcating its territory and communicating with other wolves of the pack [17].

The Bat algorithm [15] is a bio-inspired method based on the behaviour of micro-bats in their natural environment. The unique behaviour that characterize bats is their echolocation mechanism. This mechanism helps bats orient and find prey within their environment. The search strategy of bat is controlled by the pulse rate and loudness of their echolocation mechanism. Whilst the pulse rate changes to improve on better position that was previously found, the loudness indicates to each other bat that best position is accepted/found. The bat behaviour has been applied in several optimization problems to find the best optimal solution. The bat algorithm search process starts with random initialization of the population, evaluation of the new population using a fitness function and finding the best population. Unlike wolf algorithm that uses attractiveness of prey to govern its search, bat algorithm uses the pulse rate and loudness to control the search for the optimal solution.

The Particle swarm [14] is a bio-inspired method based on the swarm behaviour such as fish and bird schooling in nature. The swarm behaviour is expressed in terms of how particles adapt, exchange information and make decision on change of velocity and position within a space based on position of other neighboring particles. The advantage of swarm behaviour is that as individual particle makes a decision, it leads to an emergent behaviour. This emergent behaviour is as a result of local interaction among individual particles in a population of particles.

The novelty of this paper is the integration of RNN with LSTM, with the proposed bio-inspired/meta-heuristic search method for feature selection. The section III discusses the proposed bio-inspired search method that tune parameters unto an RNN with LSTM so as to select features.

III. Proposed Kestrel-Based Search Algorithm

The bio-inspired algorithm is based on the behaviour of Kestrel bird when hunting for a prey. The Kestrel is a kind of bird that hunts by hovering (that is flight-hunt) or from a perch. These birds are strongly territorial and hunt individually. Author of [18] has shown that during a hunt, Kestrels are imitative rather than cooperative. This suggests that Kestrels prefer not to communicate with each other but rather they imitate the behaviour of other Kestrels with better hunting technique. Authors of [19] have shown that hunting behaviour can change based on type of prey, prevailing weather conditions and energy requirements (for gliding or dive). Aside these behaviour, during hunt, Kestrels use their eyesight to watch small and agile prey within its circling radius or coverage area referred to as the visual circling radius. The minute air disturbance from flying preys, and trail

of urine and faeces from ground preys give an indication of the availability of prey. Once available prey is detected, the Kestrel positions itself to hunt. Kestrels are able to hover in changing airstream, maintain fixed forward looking position with its eye on a prey, and uses random bobbing of head to find the least distance between its position and the position of a prey. Also, the Kestrel possess an excellent ultraviolet sensitive eyesight characteristic to visually locate trails because these trails of urine and faeces of prey reflect ultra-violet light.

In hovering, Kestrel perform a wider search (global exploration) across territories within their visual circling radius, maintain a motionless position with its forward looking eye fixed on prey, detect minute air disturbance from flying prey (particularly flying insects) to best position themselves to hunt prey, and mostly move with precision through changing airstream. Kestrels are able to flap their wings and adjust their long tails to stay in a place that is referred to as a still position in changing airstream. While in perch, mostly from high fixed structures, Kestrel changes its perch every few minutes, performs a thorough search (a local exploitation using its individual hunt behaviour) of its local territory with less energy requirements than a hovering hunt, and uses its ultraviolet sensitive capabilities to detect mammals such as voles closer to a perched area. The characteristics of Kestrels are summarized as follows:

1) Soaring: gives a larger search space (global exploration) within visual coverage area.

a. Still (motionless) position with forward looking eyesight fixed on prey.

b. Encircles prey beneath with keen eyesight.

2) Perching: Each Kestrel does thorough search (local exploitation) within visual coverage area.

a. Frequent bobbing of head.

b. Attracted to prey using detected visible trail then glides to capture.

3) Imitates the behaviour of a well-adapted Kestrel.

The following assumptions are made on the characteristics of the Kestrel: the still position gives a near perfect circle, thus frequent change in a circle direction depends on position of a prey in shifting the center of its circling direction; Frequent bobbing of head gives a degree of magnified or binocular vision that helps in measuring the distance to a prey that then enables the Kestrel to move with a speed to strike; Attractiveness is proportional to light reflection; thus, the higher or longer a distance from Kestrel to the trail, the less bright a trail. This distance rule applies to both hovering height and distance away from the perch; New trails are more attractive and worth pursuing than an old trail. Thus, the trail decay or trail evaporation depends on the half-life of trail; and a Kestrel, which is not well adapted to an environment, imitates the behaviour of well-adapted kestrels.

A. Mathematical formulation on Kestrel behaviour

The proposed computational model for Kestrel's is based on the description of Kestrel's behaviour and characteristics. The following mathematical expressions depict characteristics of the Kestrel:

1) Random Encircling

Encircling is when Kestrel randomly shifts (or changes) the center of circling direction in order to recognize the current position of prey. As the prey changes its current position, Kestrel uses the encircling behaviour to randomly encircle its prey. This movement of prey determines the best possible position assumed by Kestrel. The encircling $\vec{D}$ [20] is expressed as:

$$\vec{D}=|\vec{C}*\vec{x_p}(t)-\vec{x}(t)| \tag{1}$$

Thus:
$$\vec{C}=2*\vec{r1} \tag{2}$$

Where $\vec{C}$ is the coefficient vector, $\vec{x_p}(t)$ is the position vector of the prey, and $\vec{x}(t)$ indicates the position vector of a Kestrel, *r1* and *r2* are random numbers generated between 0 and 1.

2) Current position

The current best position of Kestrel is expressed as:

$$\vec{x}(t+1)=\vec{x_p}(t)-\vec{A}*\vec{D} \tag{3}$$

Thus:
$$\vec{A}=2*\vec{z}*\vec{r2}-\vec{z} \tag{4}$$

Where $\vec{A}$ is coefficient vector, $\vec{D}$ is the encircling value obtained, $\vec{x_p}(t)$ is the position vector of the prey, $\vec{x}(t+1)$ represents the current best position of Kestrels. $\vec{z}$ represents a parameter to control the active mode with $\vec{z}_{hi}$ as the parameter for flight mode and $\vec{z}_{low}$ as the parameter for perched mode, which linearly decreases from 2 (high active mode value) to 0 (low active mode value) respectively during the iteration process. This is expressed as:

$$\vec{z}=\vec{z}_{hi}-(\vec{z}_{hi}-\vec{z}_{low})\frac{itr}{Max_{itr}} \tag{5}$$

Where *itr* is the current iteration, *Max_itr* is the total number of iterations which are performed during the search. Other Kestrels that are involved in the search update their position according to the best position of the leading Kestrel. Also, the change in position of a Kestrel in airstream depends on frequency of bobbing, attractiveness and trail evaporation. This is expressed as the following:

a) Frequency of bobbing

The frequency of bobbing *f* is used for sight distance measurement in the search space. This frequency is expressed as:

$$f^{k}_{t+1}=f_{min}+(f_{max}-f_{min})*\alpha \tag{6}$$

Where, $\alpha\in[0,1]$ is a random number to control the frequency of bobbing within a visual range. f_{max} represents the maximum frequency and f_{min} is the minimum frequency both between 1 and 0 respectively.

b) Attractiveness

Attractiveness β indicates the light reflected from a trail, which is defined by:

$$\beta(r)=\beta_o e^{-\gamma r^2} \tag{7}$$

Where $\beta_o$represents the attractiveness, γ represents variation of light intensity between [0, 1]. *r* represents the sight distance $s(x_i, x_c)$ measurement which is expressed using Minkowski distance formulation as:

$$s(x_i, x_c) = \left(\sum_{k=1}^{n} |x_{i,k} - x_{c,k}|^{\lambda}\right)^{\frac{1}{\lambda}} \quad (8)$$

Thus, $$V \leq s(x_i, x_c) \quad (9)$$

Where x_i is the current sight measurement, x_c are all potential neighboring sight measurement near x_i, n is the total number of neighboring sights, λ is the order of position being considered (that is, 2), and V is the visual range.

c) Trail evaporation

A definition of a trail is the formation and maintenance of a line [13]. In natural environment, ants use trail both to trace the path to a food source and to prevent themselves from getting stuck in a single food source. Thus, ants, using these trails, can search many food sources in a search space. As ants continue to search, trails are drawn and pheromones are deposited on a trail. This pheromone help ants to communicate with each other about the location of food sources. Therefore, other ants continuously follow this path and also deposit substances for the trail to remain fresh. Similar to ants, Kestrels use trails in search of food sources. However, these trails are rather deposited by preys which provides an indication to Kestrels on availability of food sources. The assumption is that the substances deposited by a prey is similar to pheromone deposited on ants' pheromone trail. Additionally, when the source of food depletes, Kestrels no longer follow this path that leads to the location of a prey. Consequently, the trail pheromone begins to diminish with time at an exponential rate causing trails to become old and not worth pursuing. This diminishment denotes the unstable nature of the trail substances which can be theoretically stated as: if there are *N* unstable substances in a trail with an exponential decay rate γ, then an equation can be formulated to describe how *N* substance decreases in time *t* [21]. This equation is expressed as follows:

$$\frac{dN}{dt} = -\gamma N \quad (10)$$

Since the substances are unstable, it introduces a degree of randomness in the decay process. Thus, decay rate (γ) with time (t) is re-expressed as:

$$\gamma_t = \gamma_o e^{-\varphi t} \quad (11)$$

Where γ_o is a random initial value of substance that is decreased at each iteration and where *t* is the number of iterations or time steps. $t \in [0, Max\_itr]$ where *Max_itr* is the maximum number of iterations. The decay rate $\gamma_t$at time *t* to indicate a new trail or old trail is expressed as:

$$if\ \gamma_t \rightarrow \begin{cases} \gamma_t > 1, & trail\ is\ \text{new} \\ 0, & otherwise \end{cases} \quad (12)$$

Thus, a γ_t value greater than 1 indicates that a trail is new and trail is not decayed therefore KSA explores the search area, while a $\gamma_t$value of 0 indicates that trail is old, unattractive and trail has decayed therefore KSA would not explore the search area. Again, the decay constant φ is expressed by:

$$\varphi = \frac{\phi_{max} - \phi_{min}}{t_{1/2}} \quad (13)$$

Where φ is the decay constant, $\phi_{max}$is the maximum number substances in trail, ϕ_{min} is the minimum number of substances in trail and $t_{\frac{1}{2}}$ is the half-life period of a trail. Finally, position of Kestrel is expressed by:

$$x_{i+1}^k = x_i^k + \beta_o e^{-\gamma r^2}\left(x_j - x_i^k\right) + f_i^k \quad (14)$$

Where x_{i+1}^k is the current best position of the Kestrel that represents candidate solution and $x_i^k$is the previous position of Kestrel. Where $\beta_o e^{-\gamma r^2}$represents the attractiveness as expressed in equation (7) where γ is equal to γ_t. x_j represents a Kestrel with a better position whilst f_i^k is the frequency of bobbing as expressed in equation (6).

d) Velocity

The velocity of Kestrel is updated using the expression:

$$v_{t+1}^k = v_t^k + x_t^k \quad (15)$$

Where v_{t+1}^k is the current best velocity, v_t^k represents the initial velocity, whilst x_t^k represents the current best position of Kestrel.

3) Imitative behaviour

Kestrel birds are territorial and hunt individually rather than hunt collectively. As a consequence, a model by [22] that depicts the collective behaviour of birds for feature similarity selection could not be applied. Since Kestrels are imitative, it implies that a well-adapted Kestrel would perform action appropriate to its environment, while other Kestrels that are not well-adapted imitate and remember the successful actions. The imitation behaviour reduces learning and improves upon the skills of less adapted Kestrels. The imitation behaviour is mathematically expressed and applied to select similar features into a subset. A similarity value $Sim_{value(O,T)}$ that helps with the selection of similar features is expressed by:

$$Sim_{value(O,T)} = e^{\left(\frac{-\sum |O_i - E_i|^2}{n}\right)} \quad (16)$$

Where *n* is the total number of features, $\|(O_i - E_i)\|$represents the deviation between two features where O is the observed, E_i is estimate that is the velocity of kestrel in (15). Since the deviation is calculated for each feature dimension and the possibility of large volume of features in dataset, each time a deviation is calculated only the minimum is selected (the rest of the dimension is discarded), thus, to allow the handling of different problem to different scale of dimension of data [23]. Moreover, in cases where features that were imitated are not similar (that is dissimilarity), this is calculated by:

$$dis\_sim_{value(O,T)} = 1 - Sim_{value(O,T)} \quad (17)$$

The fitness function, which is similar to fitness function formulation used by [24], to evaluate each solution is expressed in terms of classification error of the RNN and the similar value obtained from each solution. The fitness function is formulated as:

$$fitness = \rho * Sim_{value(O,T)} + dis\_sim_{value(O,T)} * \rho \quad (18)$$

Where $\rho \in (0,1)$ is a parameter that controls the chances of imitating features that are dissimilar, C_{error} is the classification error of a RNN classifier and $Sim_{value(O,T)}$ refers to the feature similarity value obtained in feature imitation.

Our method to select features uses the RNN with LSTM network (as discussed in section II) and to also make decision on classification accuracy. Authors of [24] has shown that, the less the number of features in a subset and the higher the classification accuracy, the better the solution. The proposed algorithm to implement feature selection is expressed in Table 1 as follows:

TABLE 1: PROPOSED ALGORITHMIC STRUCTURE

Set parameters Initialize population of n Kestrels using equation. Start iteration (loop until termination criterion is met) Generate new population using random encircling Compute the velocity of each kestrel using equation (15) Evaluate fitness of each solution (18) Update encircling position for each Kestrel for all i=1 to n Find the optimal features using RNN with LSTM End loop Output results

In Kestrel Search Algorithm, each kestrel referred as search agent checks the brightness of trail substances using the half-life period; random encircling of each position of a prey before moving with a velocity; imitates the velocity of another Kestrel so that each Kestrel will swarm to the best skilled search agent.

IV. EXPERIMENTAL RESULTS

A. Experimental setup

The proposed algorithmic structure was implemented in MATLAB 2018A. In each run, we performed 100 iterations to select the best/optimal parameter. The best parameter was fed into the LSTM network in which 100 epochs were performed as suggested by [25] that it guarantees optimum results on classification accuracy. To avoid the network instability, all neurons in the input to output layers on a network learned at the same rate (that is with smaller learning rate) [25]. The initial parameters for each meta-heuristic algorithm is defined as follows: KSA (Frequency of bobbing (fb=0.97); perched parameter (zmin=0.2); flight parameter (zmax=0.8); half-life parameter (half-life=0.5); dissimilarity = 0.2; similarity =0.8); PSO [14] (w=1;c1=2.5;c2=2.0); ACO [13] (α=1;ρ=0.05); BAT [15] (β=1; A=1; r=1); WSA-MP [16] v=1;pa = 0.25; α = 0.2, which were suggested by authors of the algorithms as the best parameter that guarantee an optimal solution. To test the robustness of our proposed algorithm, six benchmark datasets shown in Table 2 (from Arizona State University) were used as it represent a standard benchmark dataset with continuous data.

TABLE 2: DATASET FOR EXPERIMENT

Dataset	#of Instances	#of classes	#of features in original dataset
Carcinom	174	11	9182
Glioma	50	4	4434
Lung	203	5	3312
SMK_CAN_187	187	2	19,993
Tox_171	171	4	5748
CLL_SUB_111	111	3	11340

B. Experimental results

In order to select the best optimal solution, the study applied the concept that the higher the classification accuracy, the better the solution and hence, the less the number of features in a subset [24]. With this concept in mind, the study first applied KSA and comparative algorithms to find the best learning parameter presented in Table 3. There are ten separate runs performed on each algorithm and the best was recorded as shown in Table 3

It is observed from Table 3 that out of the six datasets, KSA has the best learning parameter (highlighted in bold) in three datasets. The learning parameter of each meta-heuristic algorithm are fed into LSTM and the classification accuracy are recorded in Table 4:

It is observed from Table 4 that the algorithm with the best parameter is not the best choice on some datasets. For instance, BAT produced the best parameter of **0.0002043** on Tox_171 dataset but produced a classification accuracy of 0.6925. It could be observed that KSA provided the highest classification accuracy on four out of six datasets. This shows that our proposed approach can explore and exploit search space efficiently and find the best results that guarantees higher classification accuracy. The results from this experiment also indicate that no single algorithm can perform better than any other. Moreover, the average classification accuracy for each algorithm when computed shows that KSA has the higher average classification accuracy of **0.7267** while PSO has least of **0.4793**. In order to select features, [24] indicated that the higher the classification accuracy, the better the solution and hence, the less the number of features in a subset. Table 5 shows the number of feature selected by each algorithm.

It is observed from Table 5 that KSA selected less number of features in **four** datasets namely **Carcinom, SMK_CAN_187, Tox_171** and **CLL_SUB_111**; PSO selected less feature in **two** datasets namely **Glioma** and **Lung**. Additionally, on average KSA selected 2422 (see table 5) features, with average accuracy of 0.7267 (see table

TABLE 3: LEARNING PARAMETERS OF META-HEURISTIC ALGORITHMS

Learning parameter	KSA	BAT	WSA-MP	ACO	PSO
Carcinom	**1.3557e-07**	1.0401e-07	3.0819e-05	8.7926e-04	0.5123
Glioma	**2.3177e-06**	3.0567e-05	1.9852e-05	9.9204e-04	0.3797
Lung	**5.1417e-06**	4.4197e-05	3.0857e-05	6.231e-04	0.3373
SMK_CAN_187	0.015064	1.338e-05	**4.7188e-05**	2.7294e-05	2.5311
Tox_171	0.16712	**0.0002043**	0.086214	0.0023152	2.2443
CLL_SUB_111	0.82116	0.075597	0.76001	**0.011556**	9.6956
Average	1.67E-01	1.26E-02	1.41E-01	2.73E-03	2.62E+00

TABLE 4: CLASSIFICATION ACCURACY OF META-HEURISTIC ALGORITHMS

Classification Accuracy	KSA	BAT	WSA -MP	ACO	PSO
Carcinom	**0.7847**	0.7806	0.6908	0.7721	0.7282
Glioma	0.7416	0.7548	0.5063	0.7484	**0.7941**
Lung	0.5754	0.5754	0.5754	0.5754	**0.7318**
SMK_CAN_187	**0.6828**	0.6759	0.6585	0.6111	0.2090
Tox_171	**0.7945**	0.6925	0.7880	0.5889	0.2127
CLL_SUB_111	**0.7811**	0.4553	0.7664	0.4259	0.2000
Average	**0.7267**	**0.6558**	**0.6642**	**0.6203**	**0.4793**

TABLE 5: FEATURE SELECTED BY EACH ALGORITHM.

Feature selected	KSA	BAT	WSA -MP	ACO	PSO
Carcinom	**1977**	2015	2839	2093	2496
Glioma	1146	1087	2189	1116	**913**
Lung	1406	1406	1406	1406	**888**
SMK_CAN_187	**6342**	6480	6828	7775	15814
Tox_171	**1181**	1768	1219	2363	4525
CLL_SUB_111	**2482**	6177	2649	6510	9072
Average	**2422**	**3156**	**2855**	**3544**	**5618**

4) and average parameter of 1.67E-01 (see table 3); while on average PSO selected 5618 (see table 5) features, with average accuracy of 0.4793 (see table 4) and average parameter of 2.62E+00 (see table 3).

The study conducted statistical test on classification accuracy to identity the best algorithm. In order not to prejudice which algorithm outperformed each other, the mean of all the algorithms were considered as equal for the statistical analysis. The Wilcoxon signed rank test which is a non-parametric statistical procedure was used because it does not make underlying assumption about the distribution of parameters and underlining dataset for the evolutionary algorithm. The advantage of Wilcoxon test is that it helps to perform pairwise comparison while not making any assumptions about the population used since Wilcoxon test can guarantee to about 95% (that is, 0.05 level of significance) of efficiency if the population is normally distributed. The results on the test statistic is shown in Table 6

TABLE 6: ALGORITHM AND P-VALUE

Algorithm	Asymp. Sig. (2-tailed) (that is, p-value)
BAT – KSA	0.225
WSAMP - KSA	0.043
ACO – KSA	0.080
PSO – KSA	0.173

Based on the results on test statistics ($p<0.05$), the following analysis can be drawn. In respect of KSA comparison with BAT, ACO and PSO, there is no statistically significant differences between the mean of classification accuracy at level of significance of 0.05. Thus, KSA is comparable to BAT, ACO and PSO algorithms. In contrast, the comparison between KSA and WSA-MP shows a statistically significant difference between the mean of classification accuracy, where $p< 0.05$ (that is, $0.043<0.05$). Thus, comparing algorithms (KSA and WSA-MP) using the Wilcoxon text show the classification accuracy of these algorithms are different.

V. CONCLUSION AND FUTURE WORK

Compared with meta-heuristic algorithms, the classification accuracy results on KSA is different from WSA-MP while the classification accuracy of KSA is comparable to ACO, BAT and PSO. The advantage of KSA is the ability to adapt to different datasets and guarantees good solutions that is comparable to other meta-heuristic search methods for feature selection.

REFERENCES

[1] Longbottom, C. and Bamforth, R., (2013), "Optimising the data warehouse." Dealing with large volumes of mixed data to give better business insights. Quocirca.

[2] Dash, M. and Liu, H. (1997), "Feature selection for classification, intelligent data analysis 1", pg 131-156.

[3] Kumar, V. and Minz, S. (2014), "Feature selection: A literature review." Smart Computing Review, vol. 4, No. 3

[4] Lin, C-J., Support vector machines: status and challenges. 2006. Available on: https://www.csie.ntu.edu.tw/~cjlin/talks/caltech.pdf

[5] Deng, Li and Yu, Dong (2013), Deep Learning: Methods and Applications. Vol. 7, Nos. 3–4 pages: 197–387.

[6] Deng, Li., Three Classes of Deep Learning Architectures and Their Applications: A Tutorial Survey. research.microsoft.com. 2013.

[7] Marcus, G. Deep Learning: A Critical Appraisal. 2018 https://arxiv.org/abs/1801.00631

[8] Patel, A. B., Nguyen, T. and Baraniuk, R. G., A Probabilistic Theory of Deep Learning. 2015.

[9] Deng, L., Three Classes of Deep Learning Architectures and Their Applications: A Tutorial Survey. 2012.

[10] LeCun, Y., Bengio, Y. and Hinton, G. 2015. Review: Deep learning. Nature. Vol. 521

[11] Li, J., Fong, S., Wong, R. K., Millham, R. and Wong, K. K. L., (2017), "Elitist binary wolf search algorithm for heuristic feature selection in high-dimensional bioinformatics datasets."

[12] Agbehadji, I. E. (2011), "Solution to the travel salesman problem, using omicron genetic algorithm. Case study: tour of national health insurance schemes in the Brong Ahafo region of Ghana." Online Master's Thesis from KNUST, Accra-Ghana.

[13] Dorigo M. and Cambardella, L. M. (1997), "Ant colony system: A cooperative learning approach to traveling salesman problem," IEEE Trans. Evol., Comput. 1 (1), pp. 53-66.

[14] Kennedy, J. and Eberhart, R. C. (1995), "Particle swarm optimization." Proc. of IEEE International Conference on Neural Networks, Piscataway, NJ. pp. 1942-1948.

[15] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), pp. 65–74, 2010

[16] Tang, R., Fong, S., Yang, X-S and Deb, S. (2012), "Wolf search algorithm with ephemeral memory."

[17] Agbehadji, I. E., Millham, R. and Fong, S. (2016), "Wolf search algorithm for numeric association rule mining." 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA 2016). Chengdu, China.

[18] Varland, D. E. (1991), "Behavior and ecology of post-fledging American Kestrels." Retrospective Theses and Dissertations Paper 9784.

[19] Vlachos, C, Bakaloudis, D., Chatzinikos, E., Papadopoulos, T. and Tsalagas, D. (2003), "Aerial hunting behaviour of the lesser Kestrel falco naumanni during the breeding season in thessaly (Greece)."

[20] Kumar, R. (2015), "Grey wolf optimizer (GWO)".

[21] Spencer, R. L. (2002), "Introduction to Matlab."

[22] Cui, X., Gao, J. and Potok, T. E. (2006), "A flocking based algorithm for document clustering analysis." 2006.

[23] Blum, A. L. and Langley, P. (1997), "Selection of relevant features and examples in machine learning." Artificial Intelligence, vol. 97, pp. 245-271.

[24] Mafarja, M. and Mirjalili, S. Whale optimization approaches for wrapper feature selection. Applied Soft Computing. 2018.

[25] Batres-Estrada, G. 2015, Deep Learning for Multivariate Financial Time Series