Vol. 11 (1) 2022

ECONOPHYSICS & SOCIOPHYSICS

other

Multidisciplinary Sciences Journal

(ESMSJ)

ESMSJ ISSN: 2247 – 2479 ISSN – L: 2247 – 2479 Vol XI, Issue 1 / 2022

Econophysics, Sociophysics & Other Multidisciplinary Sciences Journal (ESMSJ) provides a resource of the most important developments in the rapidly evolving area of Econophysics, Sociophysics & other new multidisciplinary sciences. The journal contains articles from Physics, Econophysics, Sociophysics, Demographysics, Socioeconomics, Quantum Economics, Econo-operations Research, or many other transdisciplinary, multidisciplinary and modern sciences and related fundamental methods and concepts.

SUBMIT AN ARTICLE to E-mail: gsavoiu@yahoo.com

CONTACT: +40745047085;

University of Pitesti, Adress: Str. Targul din Vale, Nr.1, Pitesti 110040, Arges,
Phone: +40 348-453100; Fax: +40 348-453123
Gheorghe Săvoiu Phone: +40745047085; E-mail: gsavoiu@yahoo.com

CONTENTS

INTER-, MULTI-, AND TRANSDISCIPLINARY MODELLING AND MODELS

Gheorghe Săvoiu

Romanian Statistical Society, Bucharest, e-mail: gsavoiu@yahoo.com

Abstract. *Perhaps the idea connected to the priority of object, method, theory or model in the modern sciences' way of thinking could have been the aim of this investigation, but inter-, multi-, and trans-dimensionality of the same model dominates the present, in science and education. Modern model is less (uni)disciplinary and dominantly inter-, multi-, and transdisciplinary. The majority of the lines in this article tries to identify an adequate answer to an ordinary investigation, everything being placed under the dazzling form of a simple question: What is the contemporary content and meaning of the word model and the real sense of the action of modelling in the modern science? The central part of this article develops three important aspects for maintaining the real development of the inter-, multi-, and trans-disciplinary modern science: i) the new paradigm of scientific model and the ascendant importance modelling in the scientific research and academic education; ii) the basic conditions of and modelling; iii) the specific architecture and paradoxes of the inter-, multi-, and transdisciplinary models and scientific modelling. Some final remarks underline the necessity of a better appreciation and implementation of modelling in education and research, and the reconfiguration of a remarkable future of the model in science.*

Keywords: *model, modelling, modelling paradigm, method, theory, scientific way of thinking, inter-, multi-, and transdisciplinarity. (uni)disciplinarity*

1. INTRODUCTION

Inter-, multi-, and transdisciplinarity have a seemingly common origin and delimit characteristic forms of the antonyms of unidisciplinarity as the knowledge acquired with the help of the unique discipline or unidisciplinarity. (Uni)disciplinarity, in its open sense, without the natural pretense of knowing everything in a limited domain is an initial natural stage of the limited scientific human knowledge and understanding. Multidisciplinarity presupposes simultaneity in the process of applying the thinking of several sciences (disciplines), interdisciplinarity designates the establishment of relations between several sciences and, finally, transdisciplinarity appears "*between disciplines, along them and above them*." [1]

(Uni)disciplinary modelling appears less and less in modern practice, respectively if we give the qualifier of (uni)disciplinary model to a model built on a dominant thinking of a discipline. The frequency of this (uni)disciplinary modelling has a minimum value as one can investigate based on a team of researchers, the real proof being its appearance in very rare case in modern scientific knowledge and research.

Multidisciplinarity of modelling presupposes that the study and research of an object of reality be realized from several points of view, descended from the multiplied thinking of several sciences at the same time. Both the modelling and the multidisciplinary research object, depending on the research result, will eventually become more enriched.

Interdisciplinarity of modelling has a diverse nuanced and purpose in direct relation to minimum two (uni)disciplinary visions, be it open, assuming phenomena, concepts and general modelling laws common to several disciplines that analyze in as varied contexts as possible, to highlight the multiple facets and possibilities of application of concepts, and laws in an increasingly varied disciplinary sphere. Interdisciplinarity favors the horizontal transfer of concepts, methods and models from one discipline to another. In interdisciplinarity, wone can detail three different degrees of their transfer, on neighboring fields, from other disciplines: i) applicative transfer; ii) epistemological transfer (cognitive); iii) transfer generating new disciplines [2] (e.g. transfer of statistical-mathematical methods in economics gave birth to econometrics, the first science created through methodological transfer, which later became a multidisciplinary type, the transfer of the econometric modelling in the space of the financial economy saturated with uncertainty generated by the theory of probabilities the financial econometric model). Interdisciplinarity of modelling is also a process of focusing or concentrating on the interstitial problems of several sciences or disciplines. The interweaving of disciplines and the coordination of research can end by adopting the same set of fundamental concepts or general methodical elements, ie by delimiting a new field of knowledge or a new discipline.

Transdisciplinarity of modelling is considered a superior form of interdisciplinarity that presupposes concepts, methods, methodology and a language that tend to become universal, being generated dynamically by the action of numerous stratifications of reality about reality etc.

The complex multidisciplinarity in modelling as a form of interpenetration of disciplines, consisting in joining certain elements of various disciplinary models, highlights their common aspects, and involves a symmetrical communication between various specialists in various disciplines, in their own axiometry.

Complex multidisciplinarity in modelling does not mean the simple juxtaposition or coexistence of models belonging to most disciplines in a single field, but it is accompanied by a transition through interdisciplinarity (e.g. a permanent transfer of informational and methodological models from discipline to discipline) to transdisciplinarity as modelling purpose, in the limiting sense of a broad dissolution of all sciences into only universal one and their models in a general and unique model, a complex fusion in a huge scientific universe or multiverse of contemporary sciences and scientific models.

Alfred Marshall inimitably described mental modelling as one that needs three great intellectual faculties: a) perception; b) motivation; c) imagination (above all). Imagination meaning is to intuit and connect the direction of events that are far away or under control. a perceptible

surface, with causes and effects, which are located at a similar distance or below the same surface. [3]

Mental modelling is the representation of our deep understanding of a portion of reality that we have realized rather theoretically and less methodically and as an experimental consequence. Any mental modelling must be flexible in the sense of reconsidering the reality studied or synthesized as a field of information extended beyond the numerically limited universe or, in other words, beyond simple mathematical modelling, becoming a filter through which to interpret reality, to it is possible to act rationally on it and, especially to select based on an optimal prognosis, the most appropriate solution or variant of action for the situation. In a sense, everything that differentiates and consolidates the idea of logical, philosophical, mathematical, physical, economic, etc. thinking can be identified and redefined one by one through the specific concept of mental modelling. [2; 3]

There are general disadvantages, respectively of most mental modelling (from the comprehensive difficulty, to the subjectivity of their interpretation, from their imperfection as a methodology, to their incompleteness as a degree of coverage of reality, etc.), but also specific (as they seem to be the names given to the components of reality, with the meaning of symbolic words, as a tool for knowing the permanent and invariable essence of things in linguistic modelling or how minimalism and non-contradiction appear in logical modelling, etc.).

Becoming famous in the vast realm of thought, the problem of the circularity of formal systems finds that the desire to express knowledge in a formal way is illusory and that it exists in main formal logic systems or related systems, relatively simple assertions or theorems that cannot be solved. In that system, the respective assertions or theorems from the analyzed model are neither provable nor unprovable, like Gödel's famous problem [4].

Contrary to the mental model, the experimental model gives priority to the idea that the reality studied as a system or as a whole, represents more than the sum of the parts, the experiment continuously offering corrections to the aggregate reality, as a support for modelling. Experimental modelling characterizes physical thinking and is much closer to nature or reality.

The solution of physical models seeks to circumvent the problem of ambiguity or contradiction by continuous experimental rectification, and happily ensures the completeness of modelling by minimalism, by returning to nature or reality, in a continuous, non-speculative but interrogative way to validate the assumptions of physical knowledge. Although apparently not dominated by details, the physical model is much more capable than other specific scientific models of reconsidering their importance through the process of validating or invalidating hypotheses with the experimental thinking's help.

2. SCIENTIFIC MODELLING AND MODELS

Science by definition is open to change and indeed science as a whole is constantly changing. The primary scientific methodology is the same and has not really needed changing (also open to change): Observe, Theorize, Test Theory with data and evidence, adjust if needed, and then let it lie out there to be tested independently by others in the future and be adjusted if needed. Methodology meaning is how to find the truth through evidence, mathematical (in modern times especially statistical) and logical argument, finally through validation or invalidation old or new theories. René Descartes was advocating in his *Discours de la méthode* that a broad interdisciplinarity seems more possible in the science's future. "*Hence we must believe that all the sciences are so interconnected, that it is much easier to study them all together than to isolate one from all the others. If, therefore, anyone wishes to search out the truth of things in serious earnest, he ought not to select one special science; for all the sciences are conjoined with each other and interdependent*…" [5]

Science as knowledge, is derived from the Latin word *scientia*, and defines a systematic ensemble of knowledge connected with nature, society, education, research and thinking. "*Scientics or scientology currently represents the science of science, an investigation into the way in which the study of nature through observation and reasoning has evolved all through several millennia of human activity. Logic is, in its capacity as a "thought that thinks of itself" the first scientific discipline achieving almost unanimous recognition."* [6]

"Mathematics has come, as a result of the studies on quantities and hierarchies, turned into theorems by means of logical derivation, to be called a science of quasi-general usefulness, yet, without physics and its necessary limits and aspect of finiteness, introduced into mathematical reasoning, the results of scientific knowledge would rather be axiomatic systems of infiniteness. Through methodically measuring the manner in which the characteristics of populations vary statistics rounds up logics, mathematics and physics, while emphasizing the importance of observation and reasoning, in much the same way as physics does, by means of experiment and simulation, in its perpetual attempt to grasp reality. And so, the broad spectrum of natural science is reached, where science describes a systematic study, or the knowledge acquired subsequent to that study conducted on nature, starting from human nature (anatomy, sociology, etc.), up to animal, and even inanimate, nature (biology, geology, etc.)." [6]

Science emerges when at least four major elements are joined together: "a *characteristic part of reality, a method for investigation, an original theory and a special model for projection."* [6] All of these elements are somehow similar with *"air, earth, water and fire of the scientific thought, combining the dangers of the new connexion between reality and theory, with idealization and pragmatism, even sometime in an excessive manner."* [3]

Who could have constituted the beginnings: *"the method, the theory or the model of thinking in the process of investigation a special reality and defining a science and his status? The abundance of data has imposed the need of clarifying the importance of the mixture of method, theory and model in the contemporary science. The synthetic quantitative determinations have often been defined as methods and they hide in their large veil of*

indicators the real meanings of qualitative information, edifying for understanding the nature, structure, territoriality, and differentiated dynamics of the specific reality. The new theories trying to understand the causes, and effects of specific phenomena, and the new tendencies, the original temporal and spatial projections have invited and still invite to reflection. Using the same way in which the small models have created new sciences, we try to understand the birth and growth of the live science's way of thinking, and their new paradigms." [1; 3] The modern science becomes a brief transformation of knowledge, from the most usual and simple access to information into a special complex way of thinking, teaching, learning and researching.

Why is the method so important? First, one can find an answer to Stefan Odobleja "*Neither the subject, nor the object are the determining factors for defining a science, but this could be only the specific method, which is indeed the essential factor generating its own paradigms*". [1;7] On the other hand, primarily nature of the reality's phenomenon reveals at least three dimensions: naturally devoid of finitude, that the first is the presence of unknown or of the limit afforded by the "observed object", the second is the limit of the observer's competence and especially the third is the limit of the method used in the characteristic analysis. Thus method is always a necessity and a limit of each science. Comprehensive knowledge of relativity or type of comprehensive analysis, limiting the presence untouchable result of "*unknown*" always gives other researchers the chance to try new solutions, because there is no specific limit in human way of thinking.

The limitation caused by observers or researchers means to understand the millennial tribute to the serenity of their exigency, and especially to reveal their own incompetence: "*I remember the days when scribes let the page empty seats*" are Confucius' words underlying the decency and modesty of any researcher or scientist… [3]

The science is also the analysis of a section of the reality as object, using methods inside a specific theory and model as an instrument for the future projections. The modern science means also a special theory able to match in a practical manner to a part of reality, and the essential instruments of forecasting and projection remain models. A scientific theory is "*a model of the universe, or a restricted part of it, and a set of rules that connect the magnitudes in the model to the observations that the researcher makes*" in the usual or day by day researchers' activity. [1]

Modern scientific models are nothing else but simple representations of complex objects, systems or events and all of these models are used as tools for understanding the nature, the population, the entire world and sometimes even universe. Models use familiar words, notions, objects to represent unfamiliar situations, events, things. Modelling is that kind of action which can help scientists to communicate their ideas, and to understand not only each other but also the processes and phenomena, helping all to make predictions. A modern model is indeed a simplified image that approximates the real complex world, but allows researchers to easily understand some of the major issues or problems and offer clarity, insight, and hopefully predictive behavior. Models are constructed from familiar objects to represent unfamiliar things. Models can help a researcher to visualize most everything, or to design impossible things in your mind, something that is really difficult to see or understand. The model essence is in its state of equilibrium between necessity and utility. A scientific model, even one empirically tested, can make use of mathematics as language, but that is not strictly necessary, just useful. All the scientific models should have the next basic features: i) all initial assumptions (hypotheses) must be scientifically sound; ii) the model's mathematical language and treatment must be self-consistent values; iii) any model must describe the largest set of the available experimental data.

3. MODELLING AND MODELS' SPECIFICITY IN MODERN SCIENCE

A scientific modelling or some realized scientific models are just simplifications that approximate the real world, but allows one researcher to easily think about problems of simplexity (as simplifying the complexity) of the same reality and get clarity, insight, and hopefully predictive behavior. So modelling and models help people in better understanding. A young researcher can learn better based on visualization, because he was born and still lives in visualization times. But, the great majority of old researchers or old teachers, and a great part of the common people cannot visualize a scientific model, even if it is an image or a solid model. A classical theory meets the conditions of optimization and adequacy to the specific reality, or the object of study of the discipline, if it satisfies at least the next major requirements (Figure no. 1):

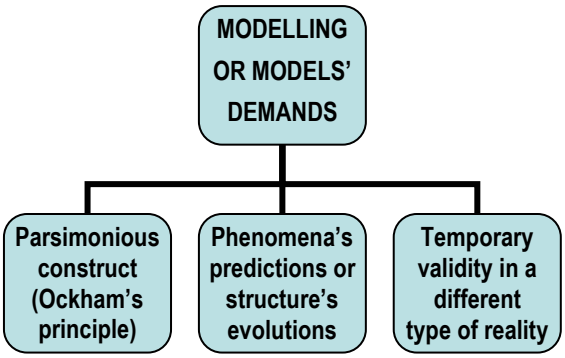


Fig. no. 1. Major Demands of classic modelling

Inter-, multi-, and trans-modelling have new requirements or mandatory needs and all of these can be synthesized as follows (Figure no. 2):

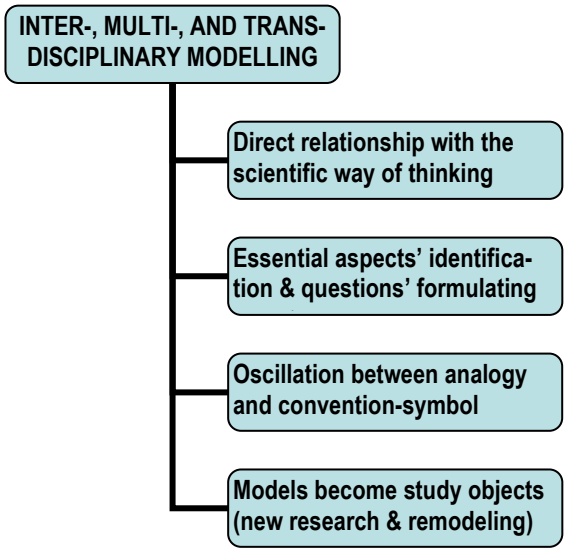


Fig. no. 2. Some essential principles of modeling

The disciplinary multiverse of today's scientific research seems to amplify the requirements of acknowledging and validation of a theory, cyclically considered as superannuated, and permanently perfectible (i.e. a theory can survive only to the extent to which its predictions are ascertained).

The theory of any scientific universe becomes, in the multiverse, a particular case of a theory much vaster in point of applicability, not yet discovered or formulated, while the new theories of the multiverse are inferences, maximized in point of coverage degree and minimized in point of mathematical and logical formulation, of the old theories, extended and selected; this fact is actually acknowledged in the very *principle of complemen-tariness* in physical thought, meaning that the old theories are particular limit cases of the new theories (where the limit, for instance in the theory of general relativity, is the speed of light, and in the theory of quantum physics – Planck's constant).

The final goal of scientific research, or even of science in general, is to provide a unique theory to supply research with a stable support in knowing and anticipating the cosmic multiverse. The multi-disciplinary model turns to account the language and methods of mathematics, testing and statistical decision, the pattern of physics in assessing reality (quantum, thermodynamic, acoustic, etc.), as well as the real variables of the specific subject to research (money flow in the economy, human behaviour in sociology, etc.). The architecture of multidisciplinary modelling capitalizes on shifting from only one science to many sciences or to a multidisciplinary model, through successive (uni)disciplinary models (improvement through imitation, analogy, and passing from one type to another).

Any inter- and multidisciplinary model can be described as an image of a specially selected part of reality, with the aid of which answers can be given to various questions, or problems belonging to an assortment of minimum two domains or fields in the area of scientific knowledge can be solved, with a certain degree of realism and with a certain limit of error. The transdisciplinary model is a result of multiple levels of reality (information theory, scientific modelling theory, systems theory etc.) [4;8]. The sad balance of the predictions made by the econometric models over the past few years, for all the modern calculation equipment added to the sophisticated classical or (uni)disciplinary models, is nothing but an additional confirmation.

All the sciences realistically recognize the impossibility of absolute modelling knowledge, but also any inter-, multi- and transdisciplinary modelling significantly increases the degree of knowledge, anticipation, structuring, etc. of that investigated reality. Emil du Bois - Reymond's famous statement "ignoramus et ignorabis" *(we don't know and we won't know ... everything* - n.a.) continuously contains a grain of truth, be it pure or only relatively. To a truth closer or farther from purity, more or less relative, (uni)disciplinary or multidisciplinary revealed, evolving from inter, to multi- or transdisciplinary, respectively one can formulate some major principles of inter-, multi- and transdisciplinary modelling and for researchers' predictions and simulations, based on these kind of interesting models [9] but, especially, paradoxically expressed. An expanded list of the major principles of inter-, multi- and transdisciplinary modelling and models must contain:

"*1. The harmony of modelling disagreements is a concord of discordances.*

2. The developmental cycle is the axis of the cyclical development.

3. The motion through an apparent state of rest, and the state of rest of the motion are the realities of all the cases of modelling. As a paraphrase to one of Schlozer's dictums, science remains history at rest, very much as history becomes science in motion.

4.The identification of the leap, or the unpredictable transformation, in the sense of the paradox of the arrow, or of the tortoise which overtakes Achilles, represents the spirit of modelling.

5. Communication, as an aim of getting out of information isolation, constitutes the message of modelling.

6. The relativity of the global interdependencies and of the local ones derives from the logic of the systems modelled, namely when the sum of the parts is greater than the whole.

7. The infinite, as part of the finite, and the finite as part of the infinite, describe the structures of modelling.

8. The finality of the inductive through deduction, and the validation of the deductive through induction bound the reasoning of those who do the modelling.

9. Knowledge is the limit to the ignorance of the act of modelling, no less than ignorance eventually becomes the result of knowledge.

10. The rebirth of theory through experiment brings about the demise of experiment in modelling.

11. The faith in critical science becomes similar to the neutrality of ignorance in the acts of modelling.

12. Coherent superposition brings together the amplitudes as limits, while incoherent superposition unites only the intensities through modelling.

13. Finding nuances is a solution of probabilistic thought, and based on the possibilities of modelling.

14. Convergence through divergence contributes to the emergence of modelling.
15. The incompleteness of completeness adds to the completeness of incompleteness in modelling.
16. The compensation of the reactions confers equilibrium to imbalance.
17. The duality of the acts of modelling is a reflex of the equivalence causes-effects.
18. A fixed multidisciplinary modelling method is no method."[6]
19. "*A model contains its own non-model, within itself or in its essence.*
20.The science of economics (financial economics) is nothing more than a long succession of econometric (financial econometric) models." [1]

All models are the expressions of some systemic approaches based on the principles of systems theory, from the principle of procedural and structural hierarchy, to the principle of dualism (dichotomy, dissonance), the principle of conservation of substance and energy, the principle of variation (general motion, oscillation, cyclicity, randomness and relativity), the principle of reactive delay or inertia, the principle of threshold value, tolerance, critical quantity, sensitivity, up to the principle of interaction.

In the case of complex real systems (political, economic, social, demographic, ecological, etc.) the modelling becomes irreplaceable, presenting two great advantages: a) pure representation of the phenomenon, process, object subject to research, without being distorted by foreign phenomena or superfluous details; b) performing experiments or performing scenarios, where this activity would be impossible due to the inaccessibility of the real object or the high cost of real action. The preservation of models or their abandonment is dictated mainly by the quality of the predictions, estimates and simulations that capitalize on them.

4. SOME FINAL REMARKS

The scientists doing modelling all day long or who work with the models the entire life will probably develop intuition. All the types of inter-, multi-, and transdisciplinary models can develop a special intuitive understanding of a system, and a good talent for estimations in a variety of normal or abnormal circumstances. An important kind of intuition comes from experience, coming from simplifying subsystems to their essential subsystems, factors, variables, structures etc., Another invaluable type of intuition is coming from accurate measurements, friendly learning instruments, simple system's governing equations, and especially from predictions, and testing all predictions. A complex model, made from hundreds or thousands of equations, variables and interactions between all the variables becomes an opportunity for a better intuition.

A memorable inter-, multi-, and transdisciplinary model must have a memorable name, a simple design, a useful algorithm to solve the real problems, a precise description of phenomenon that makes testable vision or foresight. Starting from a statistical and logical methodology, a memorable inter-, multi-, and transdisciplinary model must be also a functional instrument created to improve some explanations, to promote discussions, to make forecasts, predictions or anticipations, to offer visual images of abstract concepts etc.

There are some modelling paradoxes, coming also from a good intuition of modelling process:
1. "*Model never "proves" in the common sense.*
2. *Most of the models are wrong, but if one researcher is really lucky, he or she can find or discover a useful model.*
3. *Some models work so well that it seems silly to regard them as having no connection to reality and more than sure these models are "proved" in a weak sense*." [10]
4. One researcher can create a "model", only for manipulating it to get the needed results.
5. A model is like two edged swords: if it is properly used, it can be a boon to the mankind, but in the hands of mad or bad men, it becomes a disaster in the entire world.
6. If one researcher does not get something logical from his model, then he will term it as useless model, in spite of his useless data, structure, algorithms, variables …
7. Another model paradox is its own state of equilibrium between necessity and utility. A scientific model even an empirically tested one, can make use of mathematical language, but that is not strictly necessary, just useful.
8. Science is a systematic process of studying and understanding reality and research is also a systematic investigation, another process of experimenting to establish facts and data. The common differences between science and research are in facts, truths and errors. This aspect creates "the facts, truth and errors paradox of modelling". A model can explain facts, without finding neither the truths and nor the level of errors,
9. A model can be a substitute of reality, but it cannot be what reality really represents.
10. When model's set of assumptions or hypotheses solve two or more problems the final theory of modelling can be the result of a lucky coincidence. But when two different models make the same predictions, one researcher must think of finding a significant part of the scientific truth. This is the paradox of believing too much in coincidence (set of assumptions or hypotheses) instead of producing the same predictions.
11. Previous models have been falsified and modern science always replaced all by a new one but replaceable.
12. Always, there is a new inter-, multi-, and transdisciplinary model's paradigm that rejects the old or classic theory of (uni)disciplinary model's paradigm. There is a necessary paradigm shift.
13.The new paradox of data's simplexity is more and more important. Essential attributes of a model are coming from the observed data and from retrieval data. The more data coming from observed facts a model encapsulates, the better it is (complexity), but also more data retrieval (for usage), the more efficiently it retrieves it the better the model (simplicity).
14. The model's outputs are influenced by the presence of the researchers as observers.
15. Double liar as model's paradox is a variant of Jourdain's paradox about the opposite sides of a card. In this version of the famous paradox, any model has two opposite sides and the following words are written on

these two opposite sides of a model: A) back side – "*the sentence on the other side of this model is true*"; B) face side – "*the sentence on the other side of this model is false*." [11]

The inter-, multi-, and transdisciplinary models are the future of all modelling actions, and these modern models mean many different levels of knowledge, distinct research, specific education, another correlation between theory, practice, and technology, including morality and ethics to protect communities. Finally, one researcher can separate inter-, multi- and transdisciplinary models, putting all apart from (uni)disciplinary model by impact on prospects or foresight. Multidisciplinarity makes it easier to get better outputs. Interdisciplinarity makes the same thing, more detailed in a specific area, but relative harder than multidisciplinarity.

Transdisciplinarity gets model out of the present reality, and so the model sit around outside not in exile, but in the immediate future.

5. REFERENCES

[1] Săvoiu, G., (2013a). *Modelarea Economico-financiară: Gândirea econometrică aplicată în domeniul financiar*. [*Economic and financial modelling: Econometric thinking applied to the financial field*]. Bucharest: Editura Universitară.

[2] Săvoiu, G., (2014). The impact of inter-, trans- and multidisciplinarity on modern taxonomy of sciences, *Current Science*, vol 106(5), pp. 685-690, Available online at: http://www.currentscience.ac.in/Volumes /106/05/0685.pdf [Accessed on March 16, 2022].

[3] Săvoiu, G. (2012a). *The method, the theory and the model in the way of thinking of modern sciences*, In "Limits of knowledge society", Vol II, Epistemology and Phylosophy of Science & Economy, Editors: Sîmbotin, G. and Gherasim, O., Iaşi: Intitutul European, pp.103-122.

[4] Săvoiu, G., Iorga - Simăn, I., (2011a). *Multi-disciplinaritatea şi educaţia academică. Dialoguri argumentate*, [*Multidisciplinarity and Academic Education. Reasoned Dialogues*]. Bucharest: Editura Universitară Publishing House.

[5] Descartes, R, and Jacerme, P. (1990). *Discours de la méthode*. Paris: Pocket.

[6] Săvoiu, G. (2012b). *Econophysics: Background and Applications in Economics, Finance, and Sociophysics*, London: Publisher Academic Press Inc. Elsevier.

[7] Săvoiu, G., Iorga - Simăn, I., (2011b). *Could philosophy, biology, sociology, economics, physics, mathematics and statistics be reunited into a multidisciplinary complex learning module? Arguments pro domo for a project of a multidisciplinary educational centre Kishinev–Jass–Piteşti,* Proceedings of the International Conference „The Role of Euroregions in Sustainable Development in the Context of World Crisis: Siret-Prut-Nistru Euroregion" 2011, Iași: Ed. Tehnopres, vol IX, pages 35-42.

[8] Săvoiu, G. (2015). *Statistical thinking. The contribution of its research methods and models to modern trans-, inter- and multidisciplinarity*. Bucharest: Editura Universitară Publishing House.

[9] Săvoiu, G., (2011c). *Econometrie* [*Econometrics*, Bucharest: Editura Universitară, Publishing House

[10] Draper NR. 1987. *Empirical Model Building and Response Surfaces*. New York, NY: John Wiley & Sons.

[11] O'Connor, J.J., Robertson, E. F. (February 2005). *Philip Edward Bertrand Jourdain*. The Mac Tutor History of Mathematics archive.

A SURVEY ON DNA SEQUENCE COMPRESSION ALGORITHMS

Arunachalaprabu G. [1], Fathima Bibi K. [2]
[1]*Research Scholar in Computer Science, Thanthai Periyar Govt. Arts & Science College (A), Tiruchirappalli*
E-mail: guruarun12@gmail.com
[2]*Assistant Professor in Computer Science, Thanthai Periyar Govt. Arts & Science College (A), Tiruchirappalli*
E-mail: kfatima72@gmail.com

Abstract. *Deoxyribonucleic Acid (DNA) plays a major role in the development, growth and reproduction of all living organisms. Due to the recent development of scientific researches in biology, virology and medicine public databases are over flooded with enormous amount of DNA data. It not only faces severe challenges like storage but also restricts transmission capacity and retrieval process. Lossless DNA Compression is used to reduce the size of data, improve the capacity of storage medium and henceforth vast amount of data can be transmitted at any given time. There are many existing lossless DNA compression algorithms most of them of which are not suitable for compressing the DNA data. In addition, the development of compression algorithms that help to reduce the size of DNA data is rather a difficult task. This paper discusses the recent researches on various lossless compression algorithms. Reviews on standard algorithms are briefed. The study shows that compression of DNA sequence is vital for understanding the essential characteristics of DNA data. Two major categories namely, horizontal mode and vertical mode are focused. A comparative study about the notions of the different modes of DNA compression algorithms is analysed. To evaluate the performance of DNA compression algorithms commonly used metrics such as compression ratio, saving percentage and time taken for compression and decompression were used. An outline of some research problems that assist for further development of effective compression algorithms for DNA data and the scope for future enhancement are also discussed.*

Keywords: *Bioinformatics, Deoxyribonucleic Acid, Horizontal mode, Vertical mode, Compression Ratio.*

1. INTRODUCTION

Bioinformatics is a broad multi-disciplinary field that aims to solve biological problems using Deoxyribonucleic Acid and other related information. Deoxyribonucleic Acid, or DNA, is a long, linear vital molecule of living organisms. The primary structure of DNA molecule is a double helix strand made up of four molecules or bases namely, Adenine (A), Cytosine (C), Guanine (G), and Thymine (T).

A DNA sequence is an elongated string which comprises a set of consecutive bases (Example: chmpxx sequence - TTGAACGAGAAGCCGTATGAAATGAAAATAT).
Many researches in bioinformatics focus on the study of DNA sequences based on their functions and features. For instance, diseased DNA sequences are compared with healthiest ones to detect the major differences between them. Besides, the DNA sequences are analyzed to identify similarity between patterns. For these reasons, huge amount of DNA sequences are stored in databases. When the length of the DNA sequence increase rapidly, storage and transmission become significantly harder. In addition, it causes a major issue for many analysis tasks owing to its high memory usage and cost for computation.

Compression is an effective way for reducing the size of DNA sequence. The basic concept behind compression is to reduce the number of bits needed to store DNA sequences as they can lead to improved storage capacity and minimum network traffic. The need for compression algorithms and expertise has increased as Genome Projects resulted in an exponential growth in DNA databases. With years of research and development, there are several DNA compression algorithms available to reduce the size of DNA sequence. Compression algorithms are primarily of two types: Lossy and lossless.

- Lossy involves loss of information.
- Lossless results in no loss of information.

There are many situations that require compression where the reconstruction is to be identical to the original. In addition, there are also numerous situations in which it is not possible to relax this requirement. This opens a challenging question in research fields, such as how to reduce the size of DNA sequence without sacrificing loss of information. Therefore, lossless compression algorithms that best approximate the original dataset with reduced storage cost are likely to play an important role in DNA sequence compression.

The paper presents a general study of DNA compression algorithms that have been useful to reduce the length of the DNA sequences. Most text compression algorithms have focused on the compression of DNA sequences. However, DNA sequences often consist of many repeated and non-repeated bases. It is not easy to compress DNA sequence with good compression ratio using text compression algorithms. Some interesting compression algorithms include LZ77 (Ziv and Lempel, 77), LZ78, Prediction with Partial Match (PPM), Context Tree Weighting (CTW), GNU zip (GZip), Compress method and Bzip2. LZ77 retains a dictionary in which previously encoded input stream is stored. Sliding window method is used to examine the input stream. It is divided into two buffers: 1) Search buffer – holds recently encoded stream and 2) Look-ahead buffer – holds next segment of the stream to be encoded. At the decoding phase, a buffer is maintained equal in size to the encoder's window. A good compression ratio is achieved for many sequences. Though it requires less amount of memory more time was taken to encode the sequences [1]. LZ78 (Ziv and Lempel, 1978) uses dictionary for both encoder and decoder instead of any search buffer, look-ahead buffer or sliding window [2]. PPM method (Cleary and Witten, 1984) compresses the DNA sequences with compression ratio greater than two bits per base (bpb) [3]. CTW (Willems et al., 1995) is

suitable to compress the DNA sequences below 2 bpb [4]. GZip (Jean-loup Gailly and Mark Adler, 1992) uses adaptive Lempel-Ziv coding to compress the named files in deflate mode [5]. The performance of Compress method (Terry Welch, 1984) based on LZW coding is high with minimum memory requirements. Nevertheless, the compression ratio of compress method is significantly low [6]. In Bzip2 (Julian Seward, 1996), Burrows-Wheeler block sorting technique and Huffman coding are used to reduce the size of files [7]. However, most traditional compression algorithms have not achieved good compression results.

The paper is organized as follows: Section 2 categorizes the different DNA sequence compression techniques. The formulae of the commonly used performance metrics are shown in Section 3. Section 4 describes the recent horizontal mode DNA sequence compression algorithms. Reviews of vertical mode DNA sequence compression algorithms are discussed in Section 5. Experimental results of hybrid algorithms are shown in Section 6. Finally, Section 7 summarizes the different lossless DNA sequence compression algorithms.

2. TAXONOMY OF DNA SEQUENCE COMPRESSION TECHNIQUES

This section gives an overview of the techniques reviewed in DNA sequence compression algorithms. The classification of different DNA sequence compression algorithms are shown in Figure 1. DNA compression algorithms are classically split into two common methods: Horizontal mode and Vertical mode.

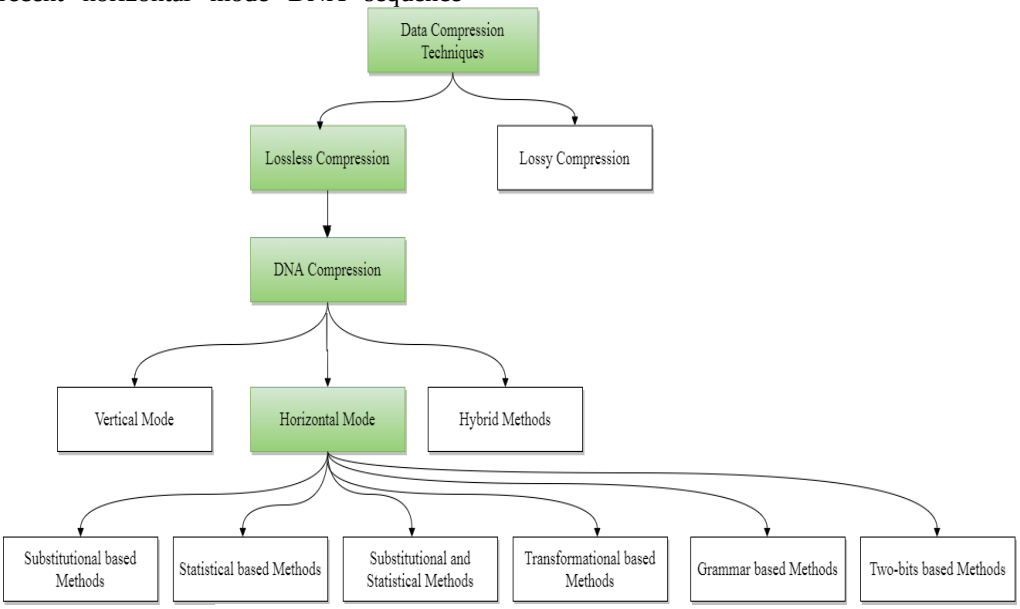


Figure 1: Taxonomy of DNA Compression Techniques

2.1 HORIZONTAL MODE

The horizontal mode compresses a sequence based on its information i.e., sequences are compressed successively. Broadly speaking, horizontal mode compression algorithms are divided into the following categories:

- Substitutional based methods – A dictionary of frequently appearing bases is maintained and when these bases appear in the sequence they are replaced by the codeword in dictionary.
- Statistical based methods – Variable size short codes are assigned to frequently appearing bases or set of bases in the sequence.
- Substitutional and Statistical based methods – Features of both substitutional and statistical methods are used to encode the sequence.
- Transformational based methods – Transformations takes place in the actual sequence and compression is applied only on the transformed sequence.
- Grammar based methods – Compresses a text string using context-free grammar. The compressed string is encoded by a symbol which in turn is converted to binary [8].
- Two-bit based methods – Unique binary bits are assigned for the bases (A = 00, C = 01, G = 10, and T = 11).

2.2 VERTICAL MODE

The vertical mode works by using the information stuck between two sequences by referring to the information contained in only one of the sequence.

3. PERFORMANCE METRICS

The effectiveness of a compression algorithm can be evaluated in various ways:

3.1 COMPRESSION RATIO (CR)

The compression ratio is the ratio between compressed file size and original file size. Compression ratio is formally expressed in bpb or bits per character (bpc).

CR = Compressed file size / Original file size

3.2 COMPRESSION FACTOR (CF)

The compression factor is the ratio between original file size and compressed file size. Compression factor is the inverse of compression ratio.

CF = Original file size / Compressed file size

3.3 SAVING PERCENTAGE (SP)

Saving percentage is the difference between original file size and compressed file size to the size of original file.

SP = (Original file size -Compressed file size) / Original file size

3.4 COMPRESSION TIME

Compression time refers to the amount of time, in milliseconds, needed to compress the file.

3.5 DECOMPRESSION TIME

Decompression time refers to the time required to decompress the compressed file to its original form. Decompression time is expressed in milliseconds.

4. HORIZONTAL MODE ALGORITHMS

With sophisticated DNA compression tasks, there is much opportunity for research and development of advanced, effectual, and scalable horizontal mode DNA compression methods in bio-informatics. Some interesting methods are:

4.1 SUBSTITUTIONAL BASED METHODS

Most compression algorithms are based on substitutional based methods. Murugan and Punitha, (2021) have designed a Pattern Matching Extended Compression Algorithm (PMECA) to compress the DNA sequence. PMECA is the extension of improved-compress algorithm [9]. First, it scans segments of the sequence and identifies identical patterns. Based on the number of bases, the patterns are stored in dictionary either in permanent or temporary manner. Matchless patterns are converted and grouped into zeros and ones. Standard datasets taken from GenBank of National Center for Biotechnology Information (NCBI) [10] was used for analysis. The algorithm resulted with a compression ratio of 91%. Simulation results have shown significant improvement of speed and reduction in file size over existing algorithms [11].

Cui et al., (2020) proposed a new approach using deep learning and arithmetic coding. In the preprocessing step, sliding window of the sequence was transformed into vectors. The local and global features are mined using Convolutional Neural Network (CNN) and Bi-directional Long Short-Term Memory Networks (BiLSTM) model. The algorithm is 3.7 times better compared to DeepDNA [12].

GeCo2 tool is an enhanced version of GeCo tool developed by Pratas et al., (2020) [13]. The genomic sequences compressed using this method are combined with cache-hash sizes, inverted repeats, interface for command line, novel pre-computed levels, and different code optimizations. The algorithm resulted with 0.2142% saving percentage when compared with GeCo.

Hui Chen (2020) suggested a genome sequence compression algorithm using entropy coding technique based on context modeling. The sequences are divided and transformed into four clusters, namely, coding sequence cluster, intron cluster, RNA cluster and residual cluster. Each set will be arranged corresponding to certain characteristics of the sequences which are encoded using entropy coding technique. The method was tested with benchmark datasets taken from US Genbank database. The algorithm resulted with an average compression ratio of 1.72 bpb [14].

Mansouri et al., (2020) described a novel lossless DNA Compression Algorithm based on Single-Block Encoding Scheme (DNAC-SBE). There are three phases namely, i) One-Bit Method Phase – position of bases with high frequencies is replaced by ones and others by zeros. ii) Single-Block Encoding Phase – encodes the generated streams and iii) Third Phase – assigns shortest codeword for each block dynamically. It is observed that DNAC-SBE has outperformed the other DNA sequence compression algorithms [15].

Shan E Zahra et al., (2019) [16] presented the Run Length Index Based Coding (RLIBC) algorithm. The basic steps are: 1) Remove all redundant DNA sequence from the input genomic dataset and store its index number 2) Perform segmentation process on each segment 3) Finally, compare each segment with index and transform the index number into binary code. When compared with other algorithms RLIBC has achieved an average compression ratio of 1.75 bpb and average compression factor of 5.7311. Data savings is 82.6% and average time taken for compression and decompression is one second.

Ayad E. Korial and Ali Kamal Taqi (2018) proposed a novel technique A2 to reduce the file size. The algorithm consists of four stages to make the substitutional model. The first stage is a modified version of Run Length Encoding which generates a symbol. The next two stages perform pre-mapping and post-mapping and the final stage develops a permutation technique using Burrows-Wheeler Transform (BWT) method. The algorithm achieved better compression ratio and saving storage space when compared with GenCompress [17]. The results of the various substitutional based compression methods are given in Table 1.

Table 1: Performance Evaluation of Substitutional Based Compression Methods

Methodology	Dataset	Performance Metrics		Drawback
		Saving Percentage	CT(sec)	
PMECA[11]	Humhdystrop	93	1.1	

	Humghcsa	90	2.1			
	Humhbb	90	2.4			
	Humhdabcd	91	1.9			
	Humhprtb	90	1.6			
		CR(bpb)				
Deep Learning Model[12]	Fish	0.70				
	Birds	0.66				
	Human	0.01				
	Ray-finned fishes	0.81				
		No. of bytes needed	CT(sec)			
GeCo2[13]	HoSa	38845642	652.4			
	GaGa	33877671	494.7			
	DaRe	11488819	198.8			
	OrSa	8646543	138.3			
	DrMe	7481093	102.4			
	EnIn	5170889	82.5			
	ScPo	2518963	34.2			
	PlFa	1925726	35.3			
	EsCo	1098552	5.1			
	HaHi	902831	4.4			
	AeCa	380115	1.9			
	HePy	375481	1.9			

		CR(bpb)	CT(sec)			
Entropy Coding Technique[14]	Chmpxx	1.5788	2.06			High compression time
	Chntxx	1.5891	1.56			
	Humhprtb	1.8532	0.57			
	Humhbb	1.8318	0.92			
	Vaccg	1.7788	3.13			
		CR(bpb)				
DNAC-SBE[15]	Chmpxx	1.604				Does not accept any other data
	Humhstrop	1.724				
	Humhbb	1.717				
	Humhcvcg	1.741				
	Mpomtcg	1.721				
	Vaccg	1.671				
	Mtpacga	1.650				
	Humhrtb	1.720				
		Percentage of Reduction				
DNAC-SBE[15]	NC_017526	23.61				Does not accept any other data
	NC_017652	22.06				
	Ce10	28.98				
	sacCer3	25.53				
	Eukaryotic	19.71				
		C-Size	C-Mem	D-Mem	CT	DT
DNAC-SBE[15]	NC_017526	0.55	36.71	64.31	1.78	1.24
	NC_017652	01.06	49.03	101.45	1.67	1.27
	sacCer3	02.48	5.09	13.54	26.57	9.71
	Ce10	19.75	68.75	212.75	73.13	45.04
	Chimpanzee	570.73	522.23	729.9	1445.15	1155.97
	Korea2009024	577.23	649.56	703.69	1479.57	1003.27
	Eukaryotic	423.32	1859.00	1349.05	1117.44	903.1
		CR(bpb)		Reduction to %		
RLIBC[16]	Humhstrop	1.400258		82.49678		

	Humhdabcd	1.414107	82.32366	
	Humhbb	1.409723	82.37846	
	Mpomtcg	1.369194	82.88507	
	Vaccg	1.406292	82.42147	
		Compression Size	Speed	CT(sec)
	gbbct45	20.9	0.453 MB/sec	196
A2[17]	gbbct108	19.1	0.455 MB/sec	195
	gbvrt9	1.70	0.456 MB/sec	17

4.2 STATISTICAL BASED METHODS

Statistical based compression methods are much familiar methods for reducing DNA sequences. Gede Eka Sulistyawan et al., (2020) have suggested a compression system which combines Burrows Wheeler Transform and Hidden Markov Model namely BWT-HMM. BWT was applied to restructure the DNA data which generates numerous redundant bases. The DNA data are segmented according to a single DNA base repeat. Re-estimation algorithm was used to reduce the storage space. The methodology was tested with DNA datasets taken from NCBI. Performance metrics such as compression ratio and time taken for computation were calculated. The proposed algorithm resulted with 4.276 bpb compression ratio with an improved mean compression ratio of 4.004 [18].

Sebastian Deorowicz (2020) introduced FQSqueezer that utilizes partial matching and dynamic Markov coder algorithms for genomic data compression. Experimentation results (Table 2) have shown that this algorithm has achieved better compression ratio for standard benchmark datasets [19].

Table 2: Performance Evaluation of Statistical Based Compression Methods

Methodology	**Dataset**	**Performance Metrics**				**Drawback**
		C-RAM	**D-RAM**	**CT (sec)**	**DT (sec)**	
FQSqueezer [19]	ERR174310_1	91.6	90.6	12728	13100	High memory usage and time
	ERR532393_1	16.4	16.4	1344	1452	
	SRR327342_1	6.7	6.6	144	145	
	SRR554369_1	6.2	6.2	68	70	
	SRR635193_1	12.1	12.1	456	462	
	SRR689233_1	11.7	11.6	406	413	
	SRR870667_1	36.4	36.1	4127	4432	
	SRR1265495_1	13.2	13.1	658	685	
	SRR1265496_1	13.0	13.0	609	652	

4.3 SUBSTITUTIONAL AND STATISTICAL BASED METHODS

It is a hybrid method that combines both substitutional and statistical approaches. Word based compression technique (Sanjeev Kumar et al., (2020)) compresses the genomic data using Modified Word Based Tag Code (MWBTC) and Delta Coding. Tests were conducted using FNA, FEN, Camera, and Eukaryotic datasets. The proposed algorithm helps to search DNA sequence devoid of decompression. When compared to LZMA and Seqcompress more than 20% to 30% better results were obtained (Table 3) [20].

Table 3: Performance Evaluation of Substitutional and Statistical Based Compression Methods

Methodology	**Dataset**	**Performance Metrics**					**Drawback**
		PCR	**C-Memory**	**D-Memory**	**CT**	**DT**	
WBCT[20]	FNA	21.52	356.89	50.78	1593	1357	Memory usage of LZMA is less compared to WBTC
	FEN	20.03	351.25	30.21	1321	1087	
	Camera	10.02	98.59	24.62	786	627	
	Eukaryotic	19.76	99.33	23.69	649	574	

4. 4 TRANSFORMATIONAL BASED METHODS

In transformational methods the DNA sequence is transformed to a specific form before compression to attain good compression ratio. Raju Bhukya (2019) developed a Differential Direct (2D) coding method based on dynamic dictionary approach. The approach works on triplets of DNA sequence bases and patterns of length multiples of three. The dictionary table of 2D coding bifurcates into two parts: i) Static part and ii) Dynamic part. The performance of the algorithm when compared with existing 2D algorithm [21] gave minimum compression ratio with reduced computational time [22].

Jothi et al., (2018) described a lossless segment compression algorithm using Lempel-Ziv Welch technique to reduce the size of DNA sequences. The architecture consists of four parts: a) Upload the DNA sequences b)

Organize the sequences c) Check relationship between two random sequences d) Compress the sequences using LZW technique. The proposed algorithm resulted with an improved compression ratio when compared to Extended ASCII algorithm, Modified RLE algorithm and COMRAD. Experimental results have shown that huge amount of time is required to arrange the sequences [23].

Shengwang Du et al., (2020) designed a compression method where the bases are converted to standard characters in first phase. The characters are compressed using LZ77 algorithm in the subsequent phase. Ten genomes of size 1 to 15M taken from NCBI database were used for testing. The performance of the proposed algorithm was measured using standard metrics such as compression ratio, compression time and decompression time. The time taken for compression and decompression is 83% and 54% respectively [24]. Table 4 gives the performance evaluation of the reviewed transformational based compression methods.

Table 4: Performance Evaluation of Transformational Based Compression Methods

Methodology	**Dataset**	**Performance Metrics**				**Drawback**
		Compressed File Size	**CR**	**CT (sec)**	**DT (sec)**	
Differential Direct Coding (2D) based on Dynamic Dictionary Approach[22]	Bacillus Subtilis	1376213	3.1061	64631	34741	Compression time is high than 2D algorithm
	Escherichia Coil K12 MG1655	1513218	3.1098	70646	37372	
	Mycoplasm Genitalium G37	185424	3.1730	8845	4267	
A compression method for DNA[24]		CR	CT (sec)	DT (sec)		Average Decompression time is minimized by 54%
	NC_017526	75.00	6.004	5.311		
	NC_002942	75.02	5.351	4.947		
	NZ_CP015934	75.05	5.985	5.073		
	NZ_CP015935	75.02	5.529	5.733		
	NZ_CP015938	75.07	5.133	5.060		
	NC_013929	75.17	9.018	17.929		
	NC_014318	75.15	9.870	15.742		
	NC_010162	75.06	12.564	25.590		

4.5 GRAMMAR BASED METHODS

In grammar based methods, context-free grammar is applied on DNA sequences. The grammars are transformed into a set of symbols and finally encoded into binary form. Diego Diaz-Dominguez and Gonzalo Navarro (2020) [25] suggested a grammar based algorithm for collection of reads to construct BWT. The collection of reads is stored as grammar to compute BWT with the support of self-indexes. The method resulted with an average compression ratio of 4.83 bpb. The study have shown that the proposed algorithm outperformed other results such as Big Repair [26], Full-text index in Minutes Space (FM-index) [27] and Run-Length FM (RLFM) [28].

4.6 TWO-BIT BASED METHODS

In two-bit based methods the bases A, C, G and T are encoded by four distinct two-bit binary values 00, 01, 10 and 11. Murugesan (2020) described a novel Codon based compression algorithm [29] based on two bit binary substitution technique. Additional dictionary is not employed to compress or decompress the genome sequence and hence additional memory is not required. Experimental results (Table 5) have shown an average compression ratio of 1.59 bpb with an average compression time of 0.18 seconds.

Table 5: Performance Evaluation of Transformational Based Compression Methods

Methodology	**Dataset**	**Performance Metrics**		**Drawback**
		CR	**CT(sec)**	
Codon Based (proposed) [29]	Humhstrop	1.55	0.095	
	Humhprtb	1.54	0.115	
	Humhbb	1.55	0.156	
	Mpomtcg	1.55	0.281	
	Vaccg	1.57	0.297	

5. VERTICAL MODE ALGORITHMS

This section reviews works that has focused on lossless DNA sequence compression algorithms based on vertical mode (Table 6). Bruno Carpentieri (2020) described a next generation sequencing data compression algorithm [30] to encode the DNA sequence using two bit encoding technique. The algorithm was tested using six DNA files namely, Lambda Virus (48,502 bytes), Homo sapiens.GRCh38.dna (3,072,712,323 bytes), SRR741411_2 (7,982,945,875 bytes), Mais (2,104,355,422 bytes), Cricetus (2,320,022,665 bytes), Pinus (20,547,720,415 bytes) to methodically demonstrate

the performance of the algorithm. The results of the proposed algorithm outperformed zip, gzip, and bzip2 algorithms.

Aníbal Guerra et al., (2020) presented UdeACompress, a referential compression algorithm to reduce the size of FASTQ files. The proposed algorithm works as follows: i) First, align the sequences to detect the most appropriate read sequence ii) Next, sort the sequence using radix sort iii) In the third phase, the sequences are encoded using binary map and instruction array techniques iv) Finally, the encoded data and unmapped reads are compressed by low level compression. The variation in file size was 14% smaller compared to the original file. Experimental results show that the time taken for execution and amount of storage was dramatically reduced and the performance of processor was improved [31].

Table 6: Performance Evaluation of Vertical Based Compression Methods

Methodology	**Dataset**	**Performance Metrics**					**Drawback**
		CR					
Proposed Algorithm [30]	LambdaVirus.fa	3.97					High Computational cost
	Homo_sapiens.GRCh38dna_sm	4.11					
	SRR741411_2	4.02					
	Mais	3.91					
	Cricetus	4.00					
	Pinus	4.01					
UdeACompress[31]		CR	CT	DT	Peak memory consumption		High memory usage and CPU requirements. Speed is sensitive.
					C	D	
	SRR1282409	7.29	2.8	10.9	10639	9691	
	SRR3141946	6.6	3.0	11.5	7578	7030	
	DRR000604	8	2.7	11.8	7162	7098	
	SRR892505	6.8	1.5	11.1	3449	3680	
	SRR892403	7.07	4.7	11.7	3414	3791	
	SRR892407	7.3	4.6	11.1	3328	3419	
SPRING[32]		Improvement					High Computational requirements
		Lossless Mode		Lossy Mode			
	Pseudomonas aeruginosa	115		62			
	Metagenomic	3206		1736			
	H.sapiens	28901		13460			
	H.sapiens	6971		5657			
	H.sapiens	25883		20316			
MZPAQ[33]		CR	CT	DT	Memory Usage		Minimum compression ratio gain
					C	D	
	SRR554369	7.04	0.98	0.91	2398.8	2383.9	
	SRR327342	8.49	1.34	1.07	2901.8	2382	
	MH0001	7.98	1.33	1.29	2691	2384.1	
	SRR1284073	3.22	0.78	0.82	2385.6	2383	
	SRR870667	6.27	0.99	0.97	4544.5	2396.3	
	ERR174310	5.00	0.83	0.99	5326.4	2383	
LFastqC[34]		CR	CT	DT	MC		Does not support color space encoding
	SRR001471	5.29	2m00s	2m16s	4		
	SRR003177	5.15	10m13s	10m43s	4		
	SRR003186	4.71	7m15s	7m59s	4		
	SRR007215	6.60	6m18s	6m08s	4		
	SRR010637	5.30	21m18s	20m59s	4		
	SRR013951	3.46	37m20s	35m27s	4		
	SRR027520_1	4.28	44m37s	48m27s	4		
	SRR027520_2	4.25	46m42s	55m49s	4		

Shubham Chandak et al., (2020) proposed a reference free compression technique for FASTQ files named SPRING. Two different modes are used precisely, lossless mode (default mode) to encode and decode FASTQ files with no loss of information and lossy mode where the arrangement of pairs and read identifiers are discarded. SPRING has achieved better results than other standard algorithms [32].

El Allali and Arshad (2019) developed a special tool called MZPAQ for compressing the genomic data in FASTQ formats. It amalgamates the features of both MFCompress and ZPAQ algorithms. The input sequence is alienated into four streams using MZPAQ. Initially, MFCompress will encode the read identifier and read sequence, next operator plus is removed and finally ZPAQ algorithm is applied. The MZPAQ achieved best compression ratio with high speed and reduced memory requirements [33].

Sultan Al YamiI and Chun-Hsi Huang (2019) proposed a lossless non-reference-based FASTQ compressor (LFastqC) which is an enhanced version of LFQC tool to decrease storage space and transmission time. The tool resulted with an enhanced compression ratio when compared with other standard algorithms. The compressor notably decreased the computation time and obtained an average compression ratio. The major drawback is that LFastqC does not support color space encoding [34].

6. HYBRID ALGORITHMS

This section discusses hybrid algorithms for DNA sequence compression (Table 7). Secure Compression Algorithm for Next Generation Sequencing (SCA-NGS) was described by Muhammad Sardaraz and Muhammad Tahir (2021). General-purpose compression library is utilized to minimize the size of quality score. The method enciphered the compressed data by applying crossover and mutation genetic algorithm concept. Results show that the proposed algorithm achieved better compression ratio of 5.08, 5.48, 5.82, 4.03, 4.65, 5.48, 5.12 and 4.19 bpb when tested with SRR801793 (2818.11), ERR022075 (11253.16), SRR125858 (52172.64), SRR611141 (1799.86), SRR489793 (13132.48), SRR935126 (10039.24), SRR003177 (1672.78) and SRR400039 (65723.77) datasets respectively [35].

Yao et al., (2021) suggested the MtHRCM and HadoopHRCM hybrid referential methods. The MtHRCM method is based on multi thread parallel technology and HadoopHRCM is implemented using distributed computing parallel technology. To assess the performance of the proposed techniques, four genomic standard datasets are chosen namely K131, YH, Huref, and HG00096 from 1000 Genome Project. The proposed methods reduced the file size from 3182 GB to 1322 MB with increased computational speed [36].

Milton Silva et al., (2020) developed a reference free and referential compression called GeCo3. The technique was applied to both multiple context model and substitution-tolerant context model of several order-depths. The algorithm mainly focuses on inputs, updates, outputs, and training process of neural networks. GeCo3 achieved better compression ratio when compared with other standard algorithms but resulted with high computational time [37].

Zeinab Nazemi Absardi and Reza Javidan (2020) proposed an innovative deep neural network based DNA sequence compression algorithm using auto encoder. Initially, the DNA sequence is preprocessed to achieve accurate results. Preprocessing is carried out in three steps. 1) Convert the characters into lowercase. 2) Delete line breaks. 3) Finally, transform non-base characters to character ‘n’. The preprocessed data is now encoded using three bit encoding scheme. A binary array is generated from the binary coded sequences. Using auto-encoder the binary array is trained and compressed. The proposed technique achieved five times better compression ratio with an improved compression accuracy of 92% [38].

Wang et al. (2018) developed DeepDNA which encompasses Convolutional Neural Network (CNN) and Long Short-Term Memory Network (LSTM) to minimize the size of genomic data. Machine learning techniques are implemented to compress the Human mitochondrial genome. The DeepDNA achieved good compression ratio of less than 0.05 bpb when compared with Gzip, MFCompress, and DMcompress [39].

Table 7: Performance Evaluation of Hybrid Compression Methods

Methodology	Dataset	Performance Metrics					Drawback
		CR	CET	CEM	DDT	DDM	
SCA-NGS [35]	SRR801793	5.09	180	1148	58	1331	Time taken for encryption is high
	ERR022075	5.48	552	1131	305	1528	
	SRR125858	4.76	2437	1638	1531	2132	
	SRR611141	4.03	102	948	36	1142	
	SRR489793	4.65	876	1536	490	1562	
	SRR935126	5.33	412	1126	193	1433	
	SRR003177	5.12	68	1638	26	1532	
		Compression Size					
MtHRCM/ HadoopHRCM [36]	chr1	108.94					
	chr2	113.25					
	chr3	98.55					
	chr4	90.54					
	chr5	81.99					
	chr6	77.91					
	chr7	73.05					

chr8	73.56	
chr9	53.31	
chr10	59.18	
chrX	48.86	
chrY	2.07	

		CR	Speed	
GeCo3[37]	HSxPT	3.65	296	Execution time is high
	HSxPA	6.57	294	
	HSxGG	4.96	293	
	GGxHS	5.81	301	
		CR	CT	
Deep Neural Network Approach [38]	KOREF_20090224	4.801	16.692	Training time was high
	KOREF_20090131	5.104	17.230	
	KOREF_20090224	4.902	27.55	
	KOREF_20090131	5.192	28.215	
	KOREF_20090224	5.003	39.002	
	KOREF_20090131	5.314	39.956	
	KOREF_20090224	5.003	42.318	
	KOREF_20090131	5.318	43.087	
		CR		
DeepDNA [39]	KF162105.1	0.01		
	MF058266.1	0.05		
	KC911416.1	0.01		
	AY339411.1	0.01		
	JQ702777.1	0.04		

7. CONCLUSION

DNA Sequence Compression is a rapidly growing and strongly related field to bioinformatics research frontiers. It is vital to study the key research issues in bioinformatics and develop new algorithms for compressing the DNA sequence for efficient analysis. The paper discusses about the classification of different lossless DNA sequence compression algorithms together with its merits and drawbacks. Some algorithms are not able to reduce the size of DNA sequences (or not achieve good compression ratio). The lossless DNA sequence compression algorithms focused include three different directions, namely, horizontal mode, vertical mode and hybrid. In each direction, different techniques are illustrated along with its experimental results such as compression ratio, time taken for compression and decompression and memory usage. Generally, horizontal mode compression techniques are applied to minimize the size of the sequences. Alternatively vertical mode compression techniques are also used to compress the sequences. Although a broad survey on the taxonomy of various lossless DNA sequence compression algorithms and their effectiveness is well beyond the scope of this survey, the results discussed here may give huge idea to readers that many remarkable works has been carried out in this analysis. Though DNA compression is highly challenging and shows potential direction, remarkable results will appear in future experiments.

8. REFERENCES

[1] Ziv, Jacob, and Lempel, A. (1977), A Universal Algorithm for Sequential Data Compression, IEEE Transactions on Information Theory, Vol. 23(3), pp. 337-343.

[2] Ziv, Jacob, and Lempel, A. (1978), Compression of Individual Sequences via Variable Rate Coding, IEEE Transactions on Information Theory, Vol. 24(5), pp. 530-536.

[3] Cleary, John, G. and Witten, H. (1984), Data Compression Using Adaptive Coding and Partial String Matching, IEEE Transactions on Communications, Vol. 32(4), pp. 396-402.

[4] Willems, F. M. J., Shtarkov, Y. M., and Tjalkens, T. J. (1995), The context tree weighting method: Basic properties, IEEE Transaction Information Theory, Vol. 41(3), pp. 653-664.

[5] Gailly, J. and Adler, M. (1992), gzip (GNU zip) compression utility. [www.gnu.org/software.gzip] website.

[6] Welch Terry, A. (1984), A technique for high performance data compression, IEEE Computer, Vol. 17(6), pp. 8-19.

[7] Julian Seward (1996), bzip2 [sourceware.org/bzip2] website.

[8] Neva Cherniavsky and Richard Ladner (2004), Grammar-based Compression of DNA Sequences, UW CSE Technical Report, pp. 1-21.

[9] Murugan, A. and Punitha, K. (2021), Pattern Matching Compression Algorithm for DNA Sequences, Proceeding of the International Conference on Sustainable Expert System, Vol. 176, pp. 387-402.

[10] Benson, D. A., Karsch-Mizrachi, I. and Lipman, D. J. (2005), GenBank, Nucleic Acids Research, Vol. 33, pp. 34-38.

[11] Murugan, A. and Punitha, K. (2021), A Pattern Matching Extended Compression Algorithm for DNA Sequences, International Journal of Computer Science and Network Security (IJCSNS), Vol. 21(8), pp. 196-202.

[12] Wenwen Cui, Zhaoyang Yu, Zhuangzhuang Liu, Gang Wang, and Xiaoguang Liu (2020), Compressing Genomic Sequences by Using Deep Learning, International Conference on Artificial Neural Networks and Machine Learning (ICANN), pp. 92-104.

[13] Diogo Pratas, Morteza Hosseini, and Armando J. Pinho (2020), GeCo2: An Optimized Tool for Lossless Compression and Analysis of DNA Sequences, International conference on Advances in Intelligent Systems and Computing, Vol. 1005, pp. 137-145.

[14] Hui Chen (2020), Application of Genome Sequence Based on Entropy Coding, International Conference on Intelligent Computing, Automation and Systems (ICICAS), pp. 156-159.

[15] Deloula Mansouri and Xiaohui Yuan (2018), One-Bit DNA Compression Algorithm, Proceedings of the International Conference on Neural Information Processing, Cambodia, pp. 376-386.

[16] Shan E Zahra, Khalid Masood and Muhammad Asif (2019), DNA Compression using an innovative Index based Coding Algorithm, IEEE 978-1-7281-4001-8/19.

[17] Ayad E. Korial and Ali Kamal Taqi (2018), Propose a Substitution Model for DNA Data Compression, International Journal of Computer Applications (0975 – 8887), Vol. 179, pp. 20-26.

[18] Gede Eka Sulistyawan, I., Achmad Arifin and Muhammad Hilman Fatoni (2020), An Adaptive BWT-HMM-based Lossless Compression System for Genomic Data, International Conference on Computer Engineering, Network and Intelligent Multimedia(CENIM 2020), pp. 429-434.

[19] Sebastian Deorowicz (2020), FQSqueezer: k-mer-based compression of sequencing data, Scientific Reports nature research, https://doi.org/10.1038/s41598-020-57452-6.

[20] Sanjeev kumar, Suneeta Agarwal and Ranvijay (2020), WBTC: A new approach for efficient storage of genomic data, International Journal of Information Technology, Springer, International Journal of Information Technology, https://doi.org/10.1007/s41870-020-00472-2.

[21] Vey, G. (2009), Differential Direct Coding: A compression algorithm for nucleotide sequence data. Database, Vol. 2009, Article ID bap013, doi:10.1093/database/bap013.

[22] Raju Bhukya (2019) Modified Direct Differential Coding Using 2D-Dynamic Dictionary for Nucleotide Sequence, Bioscience Biotechnology Research Communications, Vol. 12(4), pp. 1150-1158.

[23] Jothi, S., Chandrasekar, A., and Ranjith, R. (2018), Lossless Segment with Lempel-Ziv-Welch Compression Algorithm Based DNA Compression, Taga Journal, Swansea Printing Technology Ltd., Vol. 14, pp. 1548-1554.

[24] Shengwang Du, Junyi Li, and Naizheng Bian, A compression method for DNA, PLOS ONE, Vol. 15(11), pp. 1-8.

[25] Diego Diaz Dominguez and Gonzalo Navarro, A grammar compressor for collections of reads with applications to the construction of the BWT, IEEE DOI: 10.1109/DCC50243.2021.00016.

[26] Gagie, T., Tomohiro, I., Manzini, G., Navarro, G., Sakamoto, H., and Takabatake, Y. (2019), Rpair: Rescaling RePair with rsync, in Proc. 26th SPIRE, pp. 35-44.

[27] Ferragina, P., and Manzini, G. (2005), Indexing compressed text, J. ACM, Vol. 52(4), pp. 552-581.

[28] Gagie, T., Tomohiro, I., Manzini, G., Navarro, G., Sakamoto, H., Benkner, L., and Takabatake, Y. (2020), Practical random access to SLP-compressed texts, in Proc. 27th SPIRE, pp. 221-231.

[29] Murugesan, G. (2020), Codon Based Compression Algorithm for DNA Sequences with Two Bit Encoding, European Journal of Molecular and Clinical Medicine, ISSN 2515-8260, Vol. 07(10), pp. 33-41.

[30] Bruno Carpentieri (2020), Compression of Next-Generation Sequencing Data and of DNA Digital Files, MDPI, Algorithms, Vol. 13 (151), pp. 1-11.

[31] Anibal Guerra, Jaime Lotero and Jose Edinson Aedo (2020), Tackling the Challenges of FASTQ Referential Compression, Bioinformatics and Biology Insights, Vol. 13, pp. 1-19.

[32] Shubham Chandak, Kedar Tatwawadi, Idoia Ochoa, Mikel Hernaez, and TsachyWeissman (2018), SPRING: A next-generation compressor for FASTQ data, Oxford, Bioinformatics, Vol. 35(15), pp. 2674-2676.

[33] Achraf El Allali and Mariam Arshad (2019), MZPAQ: A FASTQ data compression tool, Source Code for Biology and Medicine, Vol. 14(3), pp. 1-13.

[34] Sultan Al Yami and Chun-Hsi Huang (2019), LFastqC: A lossless non-reference-based FASTQ compressor, PLOS ONE, pp. 1-10.

[35] Muhammad Sardaraz, and Muhammad Tahir (2021), SCA-NGS: Secure Compression algorithm for next generation sequencing data using genetic operators and block sorting, Science Progress, Vol. 104(2), pp. 1-18.

[36] Haichang Yao, Shuai Chen, Shangdong Liu, Kui Li, Yimu Ji, Guangyong Hu, and RuchuanWang (2021), Parallel compression for large collections of genomes, Concurrency Computat. Pract. Exper. John Wiley & Sons, Ltd., pp. 1-13.

[37] Milton Silva, Diogo Pratas, and Armando J. Pinho (2020), Efficient DNA sequence compression with neural networks, Gigascience, Oxford, pp. 1-15.

[38] Zeinab Nazemi Absardi and Reza Javidan (2019), A Fast Reference-Free Genome Compression Using Deep Neural Networks, Proceedings of the 2019 IEEE Conference on Big Data, Knowledge and Control Systems Engineering (BdKCSE), IEEE 978-1-7281-6481-6/19.

[39] Rongjie Wang, Tianyi Zang and Yadong Wang (2019), Human mitochondrial genome compression using machine learning techniques, Human Genomics, Vol. 13(1), pp. 1-8.

A SURVEY ON BIO-INSPIRED COMPUTING AND REVIEW OF FEATURE SELECTION BASED SWARM INTELLIGENCE

R. Senthamil Selvi[1], K. Fathima Bibi[2]

[1]Research Scholar in Computer Science, Thanthai Periyar Govt. Arts & Science College (A), Tiruchirappalli
E-mail: senthamil.behappy@gmail.com
[2]Assistant Professor in Computer Science, Thanthai Periyar Govt. Arts & Science College (A), Tiruchirappalli
E-mail: kfatima72@gmail.com

Abstract. *In recent decades, the rapid growth of database technology has led to the large-scale use of datasets. On the other hand, data mining applications work on high dimensional datasets. An important issue with applications is the term curse of dimensionality. The dimension of the data means the number of features or columns in the dataset. One of the dimensionality reduction techniques is feature selection, which means a subset of the original features. It reduces the dimensionality of data by eliminating irrelevant, redundant data. Recently, swarm intelligence techniques have gained more attention from the feature selection community because of their global search ability. In this paper, a comparative analysis, of different bio-inspired computing algorithms and recent feature selection methods based on swarm intelligence are reviewed. Furthermore, the basic operators, control parameters, variants and areas of application where these algorithms have been successfully applied. It also identifies and short listing the methodologies that are best suited for the problem. The strengths and weaknesses of the different bio-inspired algorithms are evaluated.*

Keywords: *Feature Selection, Evolutionary Computation Algorithms, Swarm Intelligence Algorithms.*

1. INTRODUCTION

Dimensionality reduction is one of the techniques used to eliminate features. Dimensionality reduction can improve the performance of machine learning algorithms and reduce computational complexity by removing irrelevant and redundant features. There are two types of approaches available in dimensionality reduction, feature selection and feature extraction. Feature extraction means creating a new set of features from the original features, whereas feature selection means a subset of the original features. Many feature selection methods use Meta heuristic optimization algorithms. It is used to find near-optimal solutions for all optimization problems. Meta heuristic algorithms are classified into Bio-stimulated algorithms, Nature-inspired algorithms, Physics-based algorithms, Evolutionary algorithms and Swarm-based algorithms.

Many bio-inspired algorithms have been employed with feature selections. A bio-inspired optimization algorithm [1] is an emerging approach; it is based on the inspiration of the biological properties of nature to develop techniques. It can be divided into 3 types as evolutionary algorithms, swarm intelligence algorithms and ecology-based algorithms. Evolutionary algorithms [2] are Darwin's theory of survival of the fittest and selection; Swarm intelligence is the behaviour of social insects such as ants, fireflies, fish, birds, bees, termites etc. Ecology-based algorithms are being used to balance the relationship between feasible and infeasible individuals.

This research work is organized as follows. Section 2 describes the taxonomy of bio-inspired computing, Section 3 discussed evolutionary based algorithms, Section 4 presents the swarm intelligence algorithms, Section 5 present the ecology based algorithms, Section 6 reviews swarm intelligence based feature selection algorithms which mainly are based on ACO, PSO, ABC, FA, GOA, WOA and GOA. The paper is concluded by Conclusions in Section 7.

2. TAXONOMY OF BIO-INSPIRED COMPUTING

This section gives an overview of the techniques reviewed in Bio-inspired algorithms. The classifications of different Bio-inspired computing algorithms are shown in Figure 1. It can be classified into 3 common algorithms like natural evolution based algorithms, swarm intelligence based algorithms and ecology based algorithms.

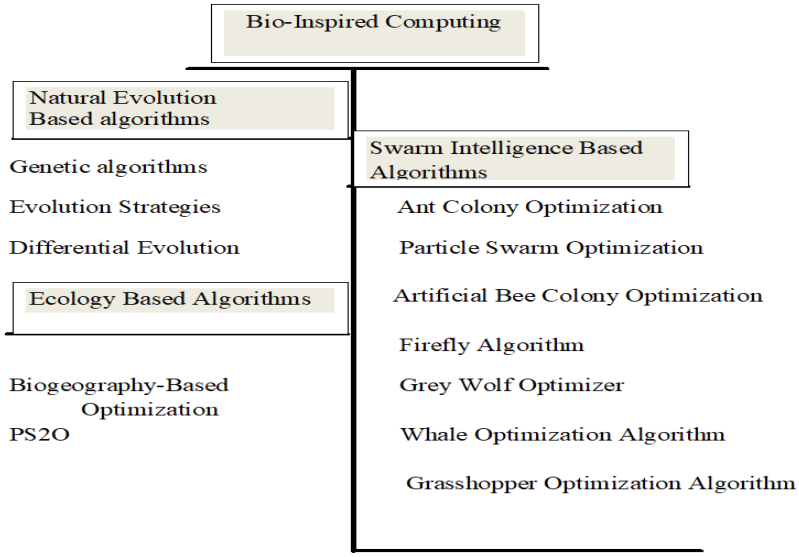


Figure 1: Taxonomy of DNA Compression Techniques

3. EVOLUTIONARY BASED ALGORITHMS

An evolutionary algorithm is a population-based Meta-heuristic algorithm inspired by nature and solves problems through the behaviors of living organisms. Evolutionary algorithms are a combination of both evolutionary computing and bio-inspired computing. The bio-inspired algorithms are based on biological evolution in nature; that is, being responsible for the design of all living beings on earth, and for the strategies they use to interact with each other.

It can be categorized into 3 types like Genetic algorithms, Evolution Strategies and Differential Evolution. These are all population based stochastic search algorithms and share a number of common features for performing with best-to-survive criteria.

3.1 GENETIC ALGORITHMS

A genetic algorithm [3] is an optimization technique based on the principles of genetics and natural selection. It was developed by John Holland and his colleagues at the University of Michigan. The phases of Genetic algorithm are Initialization of population, Fitness function, Selection, Reproduction and convergence.

3.2 EVOLUTION STRATEGIES

Evolution Strategies [4] is a type of evolutionary algorithm developed by Igno Rechenberg, Hans-Paul Schwefel and their co-workers. It is an optimization technique inspired by biological evolution and the functions may include selection, reproduction, mutation and recombination. It is commonly applied to black-box optimization problems in continuous search.

3.3 DIFFERENTIAL EVOLUTION

Differential evolution developed by Storn et al. is considered one of the population-based methods for solving complex optimization problems. Differential evolutions [5] can produce new offspring solutions through three mechanisms mutation, crossover and selection.

4. SWARM BASED ALGORITHMS

Swarm Intelligence (SI) is the concept of artificial intelligence. It was introduced by Gerardo Beni and Jing Wang. Swarm Intelligence means using the knowledge of collective objects (insects, people, etc.) together and reaching the optimal solution for a given problem. SI [6] systems are used to solve complex problems. It is the concept based on individual elements in decentralized and self-organized systems.

4.1 ANT COLONY OPTIMIZATION

Ant Colony Optimization (ACO) is one of the most successful algorithms of swarm-based algorithms. ACO was first introduced by Marco Dorigo in the 1990s. It is purely inspired by foraging behavior of ants. The ants communicate via a pheromone. The pheromone is a chemical substance that insects use to send out signals to other insects. Initially it is used to solve traveling salesman problem; later it is used for different optimization problems. In ACO [7], artificial ants are a computational agent that gives solutions to optimization problems. In the first step each ant constructs a solution; in that second step, the different ants are compared, and the last step consists of updating the pheromone levels on each stage. There are three different versions [8] of ant-system: Ant Density, Ant Quantity and Ant Cycle. Ant Density &Ant Quantity; the pheromone is updated in each movement of the ant from one place to another. Whereas Ant cycle, the pheromone is updated once all the ants have completed the tour.

4.2 PARTICLE SWARM OPTIMIZATION

Particle Swarm Optimization (PSO) is a population-based optimization technique inspired by birds of flocks and schooling in nature or insects swarming. It was proposed by Kennedy and Eberhart [9] in 1995. A collection of indi-

viduals called particles, in PSO; the particles refer to population members.

Steps in PSO

1. Generate a random population of particles: Position and Velocity.
2. Assess the position of each particle through the objective function.
3. Save each particle the best position and global best.
4. Update the velocity and the particle.
5. Go to step 2 until the stopping criteria are satisfied.

4.3 ARTIFICIAL BEE COLONY

The Artificial Bee Colony (ABC) is swarm based Meta heuristic algorithm was introduced by Karaboga [10] in 2005.ABC was inspired by the foraging behavior of honey bees. This algorithm consists of three components employed [11], onlookers' bees and scouts. The first two components are used for searching a food and third component useful for their hive. In this algorithm, the employed bees responsible for searching food using fitness values and share the information to onlooker bees. The number of employed bees or the onlooker bees is equal to the number of solutions in the population.

Steps in ABC

1. Generate food source position
2. Calculate the fitness value for each position
3. Modify neighbor positions (solutions)
4. Calculate fitness of updates position
5. Compare food positions and retain best
6. Calculate probability for positions solutions
7. Define the lowest probability for a position
8. Update position solutions
9. Go to Step 3 until the stopping criteria are satisfied.

4.4 FIREFLY ALGORITHM

Firefly Algorithm (FA) is a new Meta heuristic algorithm for optimization problems. It is inspired by the flashing behavior of fireflies. It was developed by Xin -She Yang in the year of 2008. Fireflies [12] can divide them into subgroups owed to stronger neighbor attraction over long distance attractiveness. This algorithm, randomly generated solutions called fireflies, will be assigned with a light intensity based on their performance in the objective function.

Steps in FA

1. Generate an initial population of fireflies.
2. Evaluate fitness of all fireflies from the objective function.
3. Update the light intensity fitness value of fireflies.
4. Rank the fireflies and update their position.
5. Go to Step 2 until the stopping criteria are satisfied.

4.5 GREY WOLF OPTIMIZER

Grey Wolf Optimizer (GWO) is a Meta heuristic algorithm inspired by the behavior of grey wolves in nature to hunt in a cooperative way. This algorithm developed by Seyedali Mirjalili et al. in 2014. There are four types of grey wolves [13] alpha, beta, delta and omega, where the best individual, second best individual, third best individual called α, β, δ, respectively. The remaining individuals come under the omega (ω) category.

Steps in GWO

1. Initialize the parameters like number of grey wolves, number of iterations, etc.
2. Create initial populations of grey wolves with different social hierarchy like alpha, beta, delta and omega.
3. Estimate the position of prey by alpha, beta, and delta.
4. Evaluate the position of grey wolves by the position of the grey.
5. Grade the grey wolves like the best solution called alpha, the second best solution called beta, etc.
6. Go to step 3 until the stopping criteria are satisfied.

4.6 WHALE OPTIMIZATION ALGORITHM

Whale Optimization Algorithm (WOA) is a recently developed swarm-based Meta heuristic algorithm [14] inspired by the hunting behavior of humpback whales. It was proposed by Mirjalili and Lewis in 2016. This algorithm follows bubble-net foraging behavior, which means that the whale finds its prey; it can create a bubble net along the spiral path and moves upstream to prey.

Steps in WOA

1. Initialize of search agent.
2. Calculate Fitness Value.
3. Update Whale position.
4. Apply boundary conditions and return back whales that go beyond search limits.
5. Go to step2 until the termination criteria are satisfied.

4.7 GRASSHOPPER OPTIMIZATION ALGORITHM

Grasshopper Optimization Algorithm (GOA) is a new Meta heuristic [15] algorithm and a population based algorithm inspired by the foraging and swarming behavior of grasshopper swarms in nature. It was developed by Saremi and Mirjalili, 2017. Grasshopper life cycle includes two phases called nymph and adulthood. The adulthood stage is a long range and abrupt movements where as the nymph stage is characterized by small steps and slow movements.

Steps in GOA

1. Initialize the population size, number of iterations, coefficients, and fitness function definition.
2. Assign random position of grasshoppers.
3. Evaluate the fitness of each search agent.
4. Update the position of the current search agent.
5. Check boundaries of grasshopper position.
6. Go to step3 until the termination criteria are satisfied.

5. ECOLOGY BASED ALGORITHM

The ecology-based evolutionary algorithm [16] is making attempts to balance the relationship between feasible and infeasible individuals. There can be many and complex types of interactions among the species of ecosystem. This algorithm generates considerable interest for solving real world problems. Ecology-based evolutionary algorithms are inspired by the both interspecies and intraspecies. It is more popular in solving complex multi-objective prob-

lems. It can be categorized into 2 types like Biogeography-Based Optimization and PS2O algorithm.

5.1. BIOGEOGRAPHY-BASED OPTIMIZATION

The Biogeography-Based Optimization (BBO) is used to describe the concept and models of biogeography [17]. It was developed by Simon in 2008. Biogeography is the study of the immigration and emigration of species between habitats. In BBO, each individual is termed as a habitat and has an index called the habitat suitability [18] index (HSI) to calculate its quality as a solution. It is an evolutionary algorithm that iteratively improves candidate solutions with regard to fitness functions. The operators in BBO migration and mutation are used to improve habitat solutions in the population.

Steps in BBO

1. Initialize the Habitats.
2. To compute HSI/Fitness of Habitats.
3. Perform Migration and Mutation operation.
4. Select best habitat based on HSI/Fitness value.
5. Go to step2 until the termination criteria are satisfied.

5.2. PS^2O

The PS^2O algorithm was proposed by Chen and Zhu in 2008. It is a multi-species optimizer inspired by the ideas from the co-evolution of symbiotic species in the ecosystems and it makes heterogeneous interactions between species. It is a multi-swarm approach; the interaction occurs not only between within the species but also between different species.

The following table (Table 1) summarizes the Bio-inspired computing algorithms with their control parameters, applications domain, strength and weakness of their algorithms.

Table 1: Summary of control parameters and area of application domain in Bio-inspired algorithms

Algorithm	Control Parameters	Area of Applications	Advantage	Disadvantage
Genetic Algorithm	Population size, max generation number, cross over probability, mutation probability, length of chromosome, chromosome encoding.	Robotics, Travelling and Shipment Routing	It supports multi-objective optimization.	It does not scale well with complexity.
Evolution Strategies	Population size, Maximum number of generations, Probability of crossover, Probability of mutation.	Task scheduling, car automation and Vehicle routing	Self Adaption of strategy parameters	Only applied in Continuous problems
Differential Evolution	Population size, dimension of problem, Fscale factor, probability of crossover	Image classification, filter design, chemical engineering processes and multi-objective optimization.	Reliable, accurate, robust and fast optimization technique	It is not capable of finding a new search domain.
Ant Colony Optimization	Number of ants, pheromone iterations, evaporation rate, amount of reinforcement.	Travelling Salesman Problem, Quadratic Assignment Problem, Scheduling, Vehicle routing.	Good for dynamic applications,	Convergence is guaranteed, but time to convergence is uncertain.
Particle Swarm Optimization	Number of particles, Dimension of particles, Range of particles.	Traffic Accident Forecasting, Energy-Storage Optimization, sequential ordering problem, Edge detection in noisy images, colour image segmentation.	Parallelized for concurrent processing, Efficient for global search algorithms.	Local search ability is weak.
Firefly Algorithm	Attractiveness, Randomization and absorption.	Demand Forecasting, Sensitivity Analysis.	It requires a small number of iterations.	High computational time complexity, slow convergence.
Grey Wolf Optimizer	Number of wolves, Number of iterations, problem dimension, search dimension.	Neural Network, Power System, Scheduling and Routing applications.	Simple structure so easy to implement, less storage, faster convergence.	Easy to premature, low solving accuracy.
Whale Optimization	Number of Whales,	Heterogeneous Net-	A good rate for con-	Too many parameter

Algorithm	Number of iterations, Number of variable, Random number(r).	work, Image Segmentation, Classification and Optimization.	vergence, It can handle large of decision variables, flexibility, Scalability.	tuning.
Grasshopper Optimization Algorithm	Population size, max number of iterations, coefficients, and fitness function definition.	Cloud computing, Abrupt motion tracking, Global Optimization problem.	Reasonable execution time, High accuracy, Easy to implement.	Exploiting the search space, Premature convergence in complex optimization problem.
Biogeography-Based Optimization	Number of habitats (population size), maximum migration rates, mutation rate.	Antennas and wireless communications, colour image segmentation, Satellite image classification.	An efficient algorithm for optimization. Doesn't take unnecessary computational time. Good in exploiting the solutions.	Poor in exploiting the solutions and no provision for selecting the best members from each generation.
PS^2O	Number of particles, Dimension of particles, Range of particles, Learning factors: inertia weight, maximum number of iterations.	Cooperative cognitive wireless communication, Constructing collaborative service systems (CSS).	To find a food quickly.	Take more time to allocate the food to feed.

6. FEATURE SELECTION BASED SWARM INTELLIGENCE ALGORITHMS

Feature selection is used to extract relevance data from the dataset. This section comparing the works of feature selection in swarm intelligent methods in various algorithms like ACO, PSO, ABC, FA, GWO, WOA and GOA. The following table (Table 2) consists of swarm methods comparing among the dataset, techniques, classifiers /tools, and results.

6.1 ANT COLONY ALGORITHM

Manosij Ghosh et.al (2019) proposed a filter-wrapper ACO feature selection [19] in a multi-objective manner for increasing accuracy and reducing number of features. So it can be considered as a computationally inexpensive for UCI datasets.

6.2PARTICLE SWARM OPTIMIZATION

Yu Zhou et.al (2020) proposed a model called improved discretization -based particle swarm optimization (PSO) for feature selection [20]. In this method, pre-screening process is used to reduce the size of features and then apply ranking-based cut point table are sorted the each feature and it will improve the effectiveness of the benchmark datasets.

6.3 ARTIFICIAL BEE COLONY

Artificial Bee Colony Based Feature Selection Algorithm [21] proposed by Esra Sarac Essiz et.al (2020) is effective in reducing the features and it is suitable for classification in high dimensional data. This method reduces the time without loss of accuracy in classification.

6.4 GREY WOLF OPTIMIZATION

A two-stage Improved Grey Wolf Optimization [22] called IGWO is proposed by chaonan shen et.al (2020) for feature selection on high-dimensional data. The IGWO algorithm can reduce the size of a feature, maintain high performance metrics and increase classification accuracy.

6.5 FIREFLY ALGORITHM

Sofiane-MAZA et.al (2019) proposed a firefly algorithm for feature selection [23] (FAFS) to find the best subset of the feature that gives the highest accuracy and reduces the number of features. It uses two fitness functions; they are called accuracy rate and reduction rate.

6.6 WHALE OPTIMIZATION ALGORITHM

Adel Got et.al (2021) proposed a Hybrid filter-wrapper feature selection using whale optimization algorithm [24] which uses a multi-objective approach to combine filter and wrapper for fitness functions and their experimental results show that to reduce the number of features with good classification accuracy.

6.7 GRASSHOPPER OPTIMIZATION ALGORITHM

Learning automata based improved version of Grasshopper Optimization Algorithm [25] called (LAGOA) is proposed by Chiradeep Dey et.al (2021) using two-phase mutation. The first phase reduces the number of features and the second. Phase adds relevant features which increase classification accuracy.

Table 2: Outlining the reviewed swarm intelligence based feature selection methods.

Algorithm	Algorithm/s compared with	Application	Dataset	Classifier	Accuracy
Filter-wrapper ACO Feature Selection (WFACOFS)	Feature Selection Methods	Facial Emotion recognition System.		KNN MLP	KNN MLP
			Wine		100 100
			Soy-bean-small		100 100
			Ionosphere		97.35 98.68
			Breast Cancer		99 99.67
			Monk2		88.89 87.04
			Hill-valley		55.61 64.52
			Monk1		88.89 100
			Arrhythmia		62.5 64.47
			Horse		100 100
			Madelon		100 100
Improved discretization-based Particle Swarm Optimization Feature selection(IDPSO-FS)	Potential PSO, Aadaptive Potential PSO	Mulit-objective optimization model	Adenocarcinoma	KNN	70.56
			Lymphoma		99.72
			Nic		78.74
			Colon		84.56
			Breast2		69.44
			Breast3		68.86
			Brain_tumor		89.87
			Leukemia 2		96.52
			Brain Tumor 1		87.83
			Lung cancer		93.18
Improved Grey Wolf Optimization (IGWO)	Linear forward selection, correlation based feature selection methods	Neural Network	SRBCT	MLP	100
			DLBCL		98.30
			9Tumor		63.33
			Leukemia 1		94.17
			Brain tumor 1		82.50
			Leukemia 2		97.22
			Brain tumor 2		79.17
			Prostate		94.33
			Lung cancer		98.29
			11 Tumor		93.05
Artificial Bee Colony based Feature Selection	Traditional feature selection methods like Information Gain.	Cyber bullying detection problem	Formspring	WEKA TOOL	0.72
Firefly Algorithm for Feature Selection(FAFS)	Particle Swarm Optimization for feature selection	Binary Classification		KNN NB LDA	KNN NB LDA
			Iris		96 95 98
			Wine		75.28 93.26 95.51
			Lung-Cancer		94.36 94.55 94.21
			Spambase		78.94 91.22 91.92
			Libras-Movement		98.54 96.98 95.44
			Glass		
			Segmentation		96.52 94.5 92.22
			Banknote - Authentication		96.32 94.53 93.33
			Hill-Valley		99.85 83.53 99.75
			Musk		
					96.11 93.55 94.25
					91.02 92.3 90.2
Filter-Wrapper	Single objective	Discrete problem	Breast cancer	KNN	96.70

Guided Population Archive Whale Optimization Algorithm (FW-GPAWOA)	algorithms and Multi objective algorithms		Lymphography Spect Spectf Sports articles Vehicle Whole sale customers Optical digits Letter recognition		84.61 100 84.28 84.29 74.20 92.17 94.55 93.65
Learning automata based grasshopper optimization algorithm (LAGOA)	Binary Grasshopper Optimization Algorithm (BGOA)	Disease diagnosis	Statlog (Heart) SPECTF Heart Breast Cancer (Wisconsisn) Breast Cancer (Diagnostic) Lung Cancer Hepatitis	RF	88.24 88.21 99.26 98.46 85.71 93.00

7. CONCLUSIONS

This paper provides a comprehensive survey of bio-inspired algorithms and SI based feature selection algorithms, which covers the seven most common SI algorithms: ACO, PSO, ABC, FA, GWO, WOA and GOA. In SI algorithms, the comparative analysis and categorization of different feature selection methods are evaluated. Moreover, the strengths and weaknesses of the different bio-inspired algorithms are studied. Furthermore, the algorithms in EA and SI are heuristic population-based search and it has been applied to various optimization problems in image processing, parallel computing, financial problems, forecasting problems, bio informatics etc. Nevertheless, bio-inspired algorithms are the most powerful algorithms for optimization and have a wide impact on future generation computing.

8. REFERENCES

[1] Ashraf Darwish(2018), Bio-inspired computing: Algorithms review, deep analysis, and the scope of applications, Future Computing and Informatics Journal, Volume 3, pp. 231-246, https://doi.org/10.1016/j.fcij.2018.06.001

[2] Mohamed Faiz Ahmad, Nor Ashidi Mat Isa, Wei Hong Lim, Koon Meng Ang (2022), Differential evolution: A recent review based on the state-of-the-art works, Alexandria Engineering Journal, Volume 61, Issues 5, pp. 3831-3872, https://doi.org/10.1016/j.aej.2021.09.013

[3] Sourabh Katoch, Sumit Singh Chauhan, Vijay Kumar (2021), A review on genetic algorithm past, present, and future, Multimedia Tools and Applications,80, pp. 8091–8126 https://doi.org/10.1007/s11042-020-10139-6

[4] Hans-Georg Beyer, Hans-Paul Schwefel (2002), Evolution strategies – A comprehensive introduction, Natural Computing, Volume 1, pp. 3-52, https://doi.org/10.1023/A:1015059928466

[5] Ashish Namdeo, Dileep Singh (2021), Challenges in evolutionary algorithm to find optimal parameters of SVM, Emerging Trends in Materials Science, Technology and Engineering, https://doi.org/10.1016/j.matpr.2021.03.288

[6] Suganya Selvaraj, Eumi Choi (2020), Survey of Swarm Intelligence Algorithms, International Conference on Software Engineering and Information Management, pp. 69-73, https://doi.org/10.1145/3378936.3378977

[7] Mehrdad Rostami, Kamal Berahmand, Elahe Nasiri, Saman Forouzandeh (2021), Review of swarm intelligence-based feature selection methods, Engineering Applications of Artificial Intelligence, Vol. 100, https://doi.org/10.1016/j.engappai.2021.104210

[8] Li-hua TAO, Peng-tao SHI, Jun-feng BAI (2017), Research on Parameter Optimization of ant colony algorithm based on genetic algorithm, Proceedings of the 23rd International Conference on Industrial Engineering and Engineering Management, pp. 131-136, https://doi.org/ 10.2991/978-94-6239-255-7_24

[9] Dhanalakshmi Selvarajan, Abdul Samath Abdul Jabar, Irfan Ahmed (2019), Comparative Analysis of PSO and ACO based feature selection techniques for medical data preservation, The International Arab Journal of Information Technology, Volume 16, No.4

[10] Mustafa Servet Kiran (2021), A binary artificial bee colony algorithm and its performance assessment, Expert Systems with Applications, Volume 175, https://doi.org/10.1016/j.eswa.2021.114817

[11] Dervis Karaboga, Beyza Gorkemli, Celal Ozturk (2012) , A comprehensive survey: artificial bee colony (ABC) algorithm and applications, Artificial Intelligence Review, Volume 42, https://doi.org/10.1007/s10462-012-9328-0

[12] Vijay Kumar, Dinesh Kumar (2020), A Systematic Review on Firefly Algorithm: Past, Present, and Future, Archives of Computational Methods in Engineering, Volume 28, pp. 3269-3291, https://doi.org/10.1007/s11831-020-09498-y

[13] Seyedali Mirjalili, Ibrahim Aljarah, Majdi Mafarja, Ali Asghar Heidari, Hossam Faris (2020) , Grey Wolf Optimizer: Theory, Literature Review, and Application in Computational Fluid Dynamics Problems, Nature-Inspired Optimizers, Volume 811, pp. 87-105, https://doi.org/10.1007/978-3-030-12127-3_6

[14] Nadim Rana, Muhammad Shafie Abd Latiff, Shafi'i Muhammad Abdulhamid, Haruna Chiroma (2020), Whale optimization algorithm: a systematic review of contemporary applications, modifications and developments, Neural Computing and Applications, Volume 32, pp. 16245-16277, https://doi.org/10.1007/s00521-020-04849-z

[15] Yassine Meraihi, Asma Benmessaoud Gabis, Seyedali Mirjalili, Amar Ramdane-Cherif (2021), Grasshopper Optimization Algorithm: Theory, Variants, and Applications, IEEE Access, Volume 9, pp. 50001-50024, 2021, https://doi.org/10.1109/ACCESS.2021.3067597.

[16] Ming Yuchi, Jong-Hwan Kim (2005), Ecology-inspired evolutionary algorithm using feasibility-based grouping for constrained optimization, IEEE Congress on Evolutionary Computation, Volume 2, pp. 1455-1461, https://doi.org/10.1109/CEC.2005.1554861

[17] Saman M.Almufti, Ridwan Boya Marqas, Vaman Ashqi Saeed, (2019), Taxonomy of bio-inspired optimization algorithms, Journal of Advanced Computer Science & Technology, Volume 8, pp.23-31.

[18] Dheeb Albashish, Abdelaziz I. Hammouri, Malik Braik, Jaffar Atwan (2020), Binary biogeography-based optimization based SVM-RFE for feature selection, Applied Soft Computing Journal, Volume 101, https://doi.org/10.1016/j.asoc.2020.107026

[19] Manosij Ghosh, Ritam Guha, Ram Sarkar, Ajith Abraham (2019), A wrapper-filter feature selection technique based on ant colony optimization, Neural Computing and Applications, Volume 32, pp. 7839-7857, https://doi.org/10.1007/s00521-019-04171-3

[20] Yu Zhou,Jiping Lin, Hainan Guo(2020), Feature subset selection via an improved discretization-based particle swarm optimization, Applied Soft Computing Journal,Volume 98, https://doi.org/10.1016/j.asoc.2020.106794

[21] Esra Sarac Essiz,Murat Oturakci (2020), Artificial Bee Colony-Based Feature Selection Algorithm for Cyberbullying, The Computer Journal, Volume 64,Issue 3,pp 305-313. https://doi.org/10.1093/comjnl/bxaa066

[22] Chaonan Shen, Kai Zhang (2021), Two-stage improved Grey Wolf optimization algorithm for feature selection on high-dimensional classification, Complex & Intelligent Systems, https://doi.org/10.1007/s40747-021-00452-4

[23] Sofiane Maza, Djaafar Zouache (2019), Binary Firelfy Algorithm for Feature Selection in Classfication, International Conference on Theoretical and Applicative Aspects of Computer Science, pp.1-6, https://doi.org/10.1109/ICTAACS4874.8988137

[24] Adel Got,Abdelouahab Moussaouti,Djaafar Zouache (2021). Hybrid filter-wrapper feature selection using whale optimization algorithm: A multi-objective approach, Expert System with Applications,Volume 183, https://doi.org/10.1016/j.eswa.2021.115312

[25] Chiradeep Dey, Rajarshi Bose, Kushal Kanti Ghosh (2021). LAGOA: Learning automata based grasshopper optimization algorithm for feature selection in disease datasets, Journal of Ambient Intelligence and Humanized Computing, https://doi.org/10.1007/s12652-021-03155-3

SEASONAL VARIATION IN THE AMINO ACID COMPOSITION OF THREE AIR BREATHING FISHES OF LOKTAK LAKE OF MANIPUR, INDIA

Amala_Devi_ Angom[1], Sarojnalini_Chungkham[2]

[1]Fishery Laboratory, Dept. of Life Sc. (Zoo), M.U., Canchipur-795003, India, email: angomamala2@gmail.com

[2]Fishery Laboratory, Dept. of Life Sc. (Zoo), M.U., Canchipur-795003, India, email: sarojnalini.ch@gmail.com

Abstract. *The proximate composition and amino acid profile of three important air breathing fishes Anabas testudineus, Clarias batrachus and Channa punctata of Loktak lake were determined for different seasons. Major components like moisture, protein and ash were determined using AOAC (2000) and lipid by Singh et al, (1990). Amino acids were analyzed following the methods of Ishida et al. (1981). Moisture content showed no significant variation in Clarias batrachus and Channa punctata. All the three fishes showed variation in the Protein value in all the seasons. Lipid values were recorded higher in Winter season in Clarias batrachus and Channa punctata. Higher Ash values were recorded in all the three fishes. All the three fishes had recorded higher content of histidine, lysine, serine and glutamic acid in post-monsoon season. Asparagine and Hydroxyproline were not detected. The most abundant amino acids in all the three air-breathing fishes throughout the year were Lysine, Alanine, Glycine and Phenylalanine. Essential Amino acids like Isoleucine, Leucine, Threonine and Non-essential amino acids like Aspartic acid, Arginine, Serine, Glutamine and Tyrosine are detected in lesser amount in all the three fish samples. Therefore, Anabas testudineus, Clarias batrachus and Channa punctata of Loktak lake is a good sources of nutrients.*

Keywords*: air breathing fishes, amino acids, Loktak lake, nutrients, proximate composition, seasonal variations.*

1. INTRODUCTION

Fish is an important food source globally consumed by majority of the populace. Fish plays an important role in food security in underdeveloped countries in both rural and urban areas. Fish is rich in protein, lipids, minerals, vitamins, fatty acids and amino acids, etc. The analysis of the major constituents (i.e., proximate composition) of fishes is necessary for providing information of the concentrations of protein, lipid, ash and moisture of the particular species. And, the contents of proximate composition are traditionally used as indicators of the nutritional value of fish [41].The taste of fish meat is closely related to the protein and fat content, and also the seasonal variations of these components are important determinant of both consumer choice and quality of the processed product [11]. Among the fish protein, 85-95% is digestible part which contains all dietary essential amino acid [20].

Amino acids are the major protein constituents responsible for the synthesis of most body tissues, enzymes, hormones and other metabolic molecules [28]. Fish muscle tissue is the main element for human food containing important amino acids necessary in human diet having an essential impact on growth, maintenances process, inflammation and wound healing and a unique source of physiological beneficial amino acids [14, 44, 31]. Certain amino acids like aspartic acid, glycine and glutamic acid are also known to play a key role in the process of wound healing [42]. 50-80% of the non-protein nitrogenous compounds in fish are amino acids and significant amounts of these are proline, arginine, lysine, alanine, histidine, glutamic acid and taurine [11].

Many workers has reported that nutrient composition of fish often appear to vary from season to season and the variation in the chemical composition of fish is related with their age, sex, maturity, seasons, environmental changes, etc. [20, 8, 10]. There are reports on biochemical composition of freshwater fishes from different parts of India [21, 17, 30, 32, 31] but no reports so far on seasonal variation of nutritional properties of air breathing fishes of Loktak lake of Manipur. The Present study is on the seasonal variation in the proximate and amino acid composition of three air breathing fishes; *Anabas testudineus*, *Clarias batrachus* and *Channa punctata* of Loktak lake and the findings could be helpful to nutritionists, dieticians, researchers, fish farmer, etc for future references.

2. MATERIALS & METHODS

2.1 STUDY SITE AND SAMPLE COLLECTION

The live fish samples were collected from the Loktak lake with the help of local fisherman. The samples were collected during the Pre-monsoon (March–May) Monsoon (May – August), Post-monsoon (September – November) and Winter (December – February) seasons of 2017 and 2018. The sample fish were brought to the Fishery Laboratory, Life Sciences Department (Zoology), Manipur University. Six numbers of *Anabas testudineus*, *Clarias batrachus* and *Channa punctata* of similar weight of about 70-80g, 125-155g and 80-90g and standard length of about 16-18cm, 24.5-27cm and 16-18.5cm respectively were used for different seasons. The fishes were washed thoroughly with running tap water, beheaded, eviscerated and the edible muscle parts were taken for various analyses.

2.2 PROXIMATE COMPOSITION ANALYSIS

Moisture content was determined by hot air oven method [5] at 60˚C till a constant weight is obtained. Total Nitrogen content was determined by using modified Micro-Kjeldahl's method [5]. The samples were subjected to digestion,

nesslerization and finally absorbance were measured in 440 nm by using Eppendorf BioSpectrometer. Total protein was obtained by multiplying the nitrogen value with 6.25 [29]. Total lipid content was determined as per the modified method of [39] by extraction with chloroform and Methanol in the ratio of 2:1. Ash content was determined by igniting the moisture free sample at 550˚C in a Muffle furnace for about 2-3 hours to obtained carbon free white ash as described by [5].

2.3 AMINO ACID ANALYSIS

Amino acid analysis was done following the methods of [19]. 100mg of sample was taken in a test tube and 6N HCl was added to it. The tube was filled with nitrogen gas and sealed and allowed to digest at 110˚C for 24 hours. The test tubes were cooled and then digested samples were filtered. The filtrates were evaporated to dryness by using vacuum evaporator. 10ml of Millipore water was added and evaporation was continued until the samples were acid free. Then, the acid free samples, containing free amino acids were dissolved in 10 ml of 0.05M HCl.

The digested samples were filtered by passing through Whatman filter paper no. 42 (0.45µm pore size). 20µl of this filtrate was injected through the sample loop of HPLC, fitted with a packed column (C18 reverse phase; ISC-07/51504-Na) of length 19 cm and diameter 5 mm. oven temperatutre was adjusted for 60˚C. The amino acids were detected by Spectroflurometer after post column derivatization with ophthaldehyde at the wavelength of 338 nm.

2.4 STATISTICAL ANALYSIS

The samples were analyzed using one way-ANOVA and the significant mean were compared by Duncan's multiple range tests ($P<0.05$). Data were analzed using SPSS package (version 16.0) [40].

3. RESULTS AND DISCUSSIONS

The proximate composition of three air breathing fishes *Anabas testudineus*, *Clarias batrachus* and *Channa punctata* are shown in Fig 1, 2 and 3.

In *Anabas testudineus*, moisture content ranges from 77.60±0.13−78.48±0.11% and showed no significant variation among Pre-monsoon, Monsoon and Winter season (Fig.1). In *Clarias batrachus*, moisture values ranges from 78.88±0.32−79.15±0.23% and no significant variation in moisture content was observed in all the season (Fig.2).

In *Channa punctata*, moisture content showed no significant variation in all the seasons and moisture value ranges from 80.26±0.16−80.71±0.23% (Fig.3). The moisture content in the fish muscle of all the three air-breathing fishes were within the acceptable range (60−80%). The reason might be due to the stable water levels in the environment from where the fishes were collected. [33], reported that the moisture content of some freshwater fishes of Manipur were in the range of 71.00−80.00%. [21] also reported that the moisture content of 25 different freshwater fishes were in the range of 73−82%. Among the three fish sample *Channa punctata* was recorded with highest moisture content which implies the flesh of *Channa punctata* has high water holding capacity. [7] stated that high moisture content could play important roles in metabolic reactions and also help in easily solubilize certain elements. The difference in moisture value of three air-breathing fishes might be due to difference in species, sex, feeding habits, spawning period, metabolic activities, etc.

In *Anabas testudineus*, protein content was recorded in the range of 6.43±0.45% to 12.86±0.57% and higher protein content was found in Post-monsoon season followed by Winter season. In *Clarias batrachus*, the protein content ranges from 8.06±1.24% to 15.83±0.14% and was recorded higher in Pre-monsoon season and lower in Winter season. In *Channa punctata*, the protein content ranges from 5.31±0.30% to 11.16±1.07% and was observed higher in Post-monsoon season and lower in Monsoon season. The difference in the value of protein might be attributed to difference in food

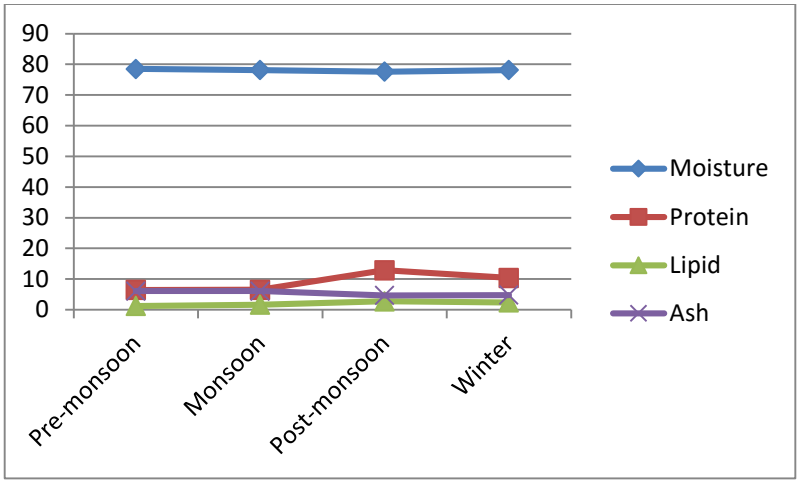


Fig:1: Moisture, Protein, Lipid and Ash content (%) of *Anabas testudineus* for four different seasons.

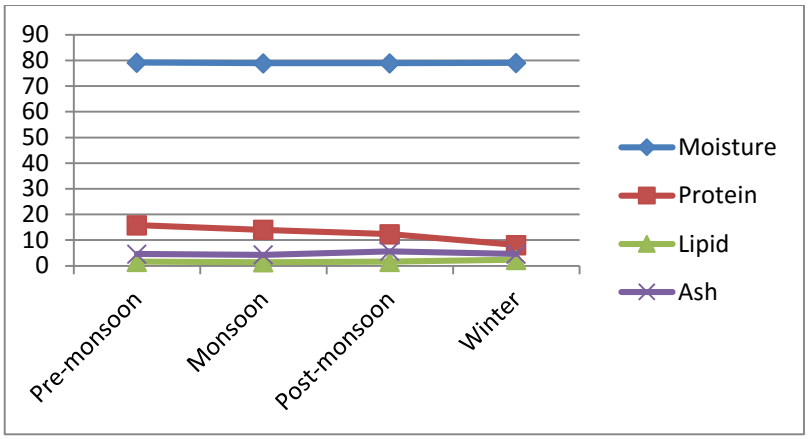


Fig.2: Moisture, Protein, Lipid and Ash content (%) of *Clarias batrachus* for four different seasons

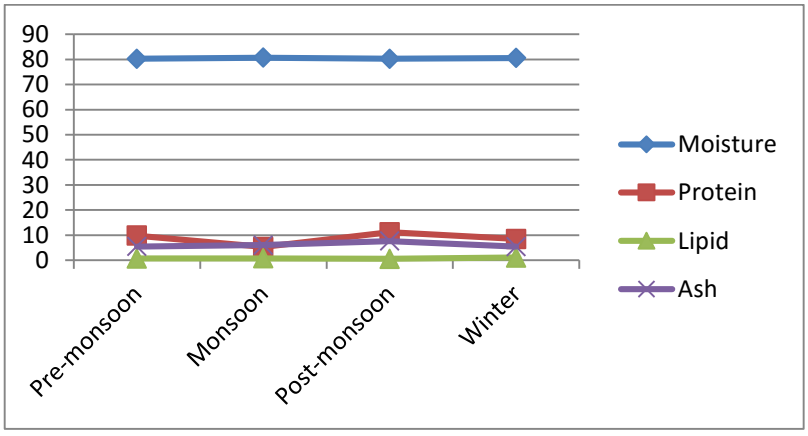


Fig.3: Moisture, Protein, Lipid and Ash content (%) of *Channa punctata* for four different seasons.

intake and availability of food, temperature and maturity. [16] also stated that, the variability in the content of protein in fish muscle depends on the abundance and availability of fish food and there is an opposite relationship among protein and moisture content. According to [43]; [37], the decrease in the muscle protein during Winter season might be due to cold climatic condition of the study area and less food intake due to shortening of the daytime.

In *Anabas testudineus*, lipid value was recorded in the ranges from 1.21±0.20% to 2.75±0.15% and higher lipid content was recorded in Post-monsoon followed by Winter

season. In *Clarias batrachus,* the lipid content showed no significant variation among Pre-monsoon, Monsoon season and Post-monsoon season, and higher lipid content was found in Winter season (Fig 2). In *Channa punctata*, the lipid content showed no significant variation among Pre-monsoon, Monsoon season and Post-monsoon season and higher lipid content was found in Winter season (Fig 3). Variation in the lipid values might be due to poor storage mechanism and the used of the fat reserves during spawning activities. According to the classification given by [1], *Anabas testudineus* and *Clarias batrachus* can be grouped into low fat fish (2-4%) and *Channa punctata* into lean fish (<2%).

In *Anabas testudineus*, ash values ranges from 4.7±0.28% to 6.1±0.07% and higher ash content was recorded in Monsoon season. In *Clarias batrachus*, ash values ranges from 4.3±0.42% to 5.65±0.35% and higher value was observed in Post-monsoon season. The ash content of *Channa punctata* showed significant variation among Monsoon, Post-monsoon and Winter, and higher ash content was observed in Post-monsoon season (Fig 3). Higher ash content in the three fishes shows that they are a good source of minerals like Na, K, Ca, Zn, Fe, etc. Higher ash content observed during Monsoon and Post-monsoon season of the year might attribute towards a higher mineral metabolism during this period.

Amino acid content of the three air breathing fish *Anabas testudineus*, *Clarias batrachus* and *Channa punctata* are shown in Table 1, 2 and 3 respectively. Significant variations were observed in all the seasons among the three samples and within the same species also.

In *Anabas testudineus*, out of 21 amino acids analyzed, 7 Essential amino acids and 8 Non-essential amino acids were detected (Table 1) and amino acids were abundant in Post-monsoon and less in Pre-monsoon season. In *Clarias batrachus*, out of 21 amino acid analyzed, 6 Essential amino acids and 8 Non-essential amino acids were detected (Table 2) and amino acids were abundant in Post-monsoon and less in Pre-monsoon season and in *Channa punctata*, out of 21 amino acids analyzed 6 Essential amino acids and 8 Non-essential amino acids were detected (Table 3) and amino acids were abundant in Winter and less in Post-monsoon season. Most of the EAAs and NAAs of the three fishes were found higher in the Post-monsoon season than in the other three seasons which is in agreement with the study of [12]. However, it was not similar with that of [32] in *Anabas testudineus* and *Clarias batrachus* for Pre-monsoon and Post-monsoon seasons.

Amino acids content in all the three fishes were detected in small amount as compare to the amino acids content reported by other workers: [38, 13, 22, 12].

Minimal quantity of amino acids were also detected by [25 in the muscle tissues of *L. niloticus, B. bayad, O. niloticus, S. schall and Tetraodon lineatus*; [23]; [32] in *Monopterus cuchia* ; [11].

All the three fishes had recorded higher content of histidine, lysine, serine and glutamic acid in post-monsoon season. Presence of high level of histidine will contribute to better taste [17]. Lysine helps the body to absorb calcium and plays an important role in the formation of collagen. Lysine is an essential amino acid which is extensively required for optimal growth and its deficiency leads to immunodeficiency [9]. Serine is important in metabolism and participates in the biosynthesis of proteins, cysteine, glycine, purines and pyrimidines. It is also being used for the treatment of schizophrenia [27]. Glutamic acid plays an important role in amino acid metabolism because of its role in transamination reactions and is necessary for the synthesis of glutathione which are required for removal of highly toxic peroxides and the polyglutamate folate cofactors [27]. Freshwater species may be a good source of glutamic acid [34].

The most abundant amino acids in all the three air breathing fishes throughout the year were lysine, alanine, glycine and phenylalanine. Some amino acids like glutamic acid, glycine and alanine are related to flavor of fish [11]. Alanine is involved in sugar and acid metabolism, increase immunity and provides energy for muscles tissue, brain and the central nervous system. Glycine is an important component of the human skin collagen that combines with aspartic and glutamic acids to form a

Table 1. Amino acids profile (mg/100g) of *Anabas testudineus* for different seasons.

Particular	Pre-monsoon	Monsoon	Post-monsoon	Winter
Essential Amino Acids				
Histidine	0.020±0.000[a]	0.046±0.002[b]	**0.132±0.001[d]**	0.082±0.000[c]
Isoleucine	0.039±0.000[b]	**0.040±0.002[b]**	0.038±0.000	0.035±0.000[a]
Leucine	0.189±0.000[c]	0.181±0.002[b]	**0.201±0.000[d]**	0.146±0.002[a]
Lysine	0.130±0.000[a]	0.168±0.002[b]	**0.336±0.001[d]**	0.207±0.000[c]
Metheonine	ND	ND	ND	ND
Phenylalanine	**0.519±0.001[d]**	0.033±0.001[a]	0.054±0.000[c]	0.038±0.000[b]
Threonine	0.099±0.000[b]	0.100±0.001[b]	0.082±0.000[a]	**0.142±0.002[c]**
Valine	0.092±0.000[a]	ND	**0.570±0.001[b]**	ND
Tryptophan	ND	ND	ND	ND

Non-essential amino acids				
Alanine	0.761±0.000[a]	0.973±0.002[c]	**1.038±0.001[d]**	0.924±0.000[b]
Aspartic acid	0.212±0.00[a]	0.222±0.002[b]	**0.334±0.000[d]**	0.256±0.000[c]
Asparagine	ND	ND	ND	ND
Arginine	0.146±0.002[a]	0.218±0.000[c]	**0.289±0.000[d]**	0.176±0.000[b]
Serine	0.147±0.000[a]	0.219±0.002[b]	**0.296±0.000[d]**	0.237±0.001[c]
Glutamic acid	0.396±0.002[c]	0.374±0.001[b]	**0.512±0.000[d]**	0.369±0.000[a]
Glutamine	0.072±0.000[b]	0.051±0.000[a]	**0.116±0.002[d]**	0.097±0.001[c]
Glycine	0.341±0.001[a]	0.428±0.002[b]	**0.798±0.000[d]**	0.630±0.000[c]
Tyrosine	0.112±0.000[a]	0.111±0.000[a]	**0.313±0.002[c]**	0.131±0.001[b]
Cysteine	ND	ND	ND	ND
Hydroxyproloine	ND	ND	ND	ND

ND- Not Detected
Values are mean of three replicates.
Means (±SD) followed the same letter are not significantly different (P≤0.05).

polypeptide, which is responsible for tissue growth and the healing of wounds [3]. Phenylalanine is the precursor of some hormones and the pigment melanin in hair, eyes and tanned skin [6]. [11], reported that the most abundant amino acids in his investigation in *Upeneus moluccensis* were lysine, leucine, aspartic acid, glutamic acid, alanine and glycine. [2] reported that lysine, arginine, leucine, glutamic acid, aspartic acid and glycine were the most abundant amino acid in his study in smoke fishes. [30], reported that lysine, leucine, threonine, phenylalanine, aspartic acid, glutamic acid and alanine were predominant in his investigation in *Tenualosa ilisha.* In the study conducted by [22] in *Mugil cephalus*, lysine, leucine, arginine, leucine, aspartic acid, glutamic acid and alanine were reported as the most abundant amino acids. [26], reported that the most abundant amino acids in four commercial Niles fishes in Sudan were glutamic acid, leucine, aspartic acid, alanine and glycine. As reported by [36, 11], fish body composition especially of fatty acids and amino acids were varied among the fishes and within the same species and the variability may depend on such factors as food availability, fishing location, fish size, maturity stage and biological variations, seasonal conditions, age and spawning season.

Asparagine and hydroxyproline were not detected in the present investigation. Absence of some amino acids like asparagines, glutamine and tryptophan could be as a result of acid hydrolysis [4] or oxidation as it destroys or chemically modifies asparagines, glutamine and tryptophan residues in protein while asparagines and glutamines are converted to their corresponding acids (aspartic and glutamic acids) and are quantified with them, tryptophan is completely destroyed or by the mere absence of these amino acids [34]. Protein

Table 2. Amino acids profile (mg/100g) of *Clarias batrachus* for different seasons.

Particular	Pre-monsoon	Monsoon	Post-monsoon	Winter
Essential Amino Acids				
Histidine	0.008±0.000[a]	0.097±0.002[c]	**0.122±0.002[d]**	0.091±0.002[b]
Isoleucine	ND	0.022±0.001[b]	0.017±0.002[a]	**0.029±0.002[c]**
Leucine	0.066±0.001[a]	0.111±0.002[b]	**0.148±0.002[d]**	0.121±0.000[c]
Lysine	0.021±0.000[a]	0.263±0.002[b]	**0.397±0.002[d]**	0.377±0.001[c]
Metheonine	ND	ND	ND	ND
Phenylalanine	ND	0.028±0.002[a]	**0.054±0.001[c]**	0.048±0.002[b]
Threonine	0.004±0.001[a]	**0.173±0.002[d]**	0.163±0.001[c]	0.131±0.002[b]
Valine	ND	ND	ND	ND
Tryptophan	ND	ND	ND	ND
Non-essential amino acids				
Alanine	0.038± 0.002[a]	**6.676±0.000[d]**	0.581±0.000[b]	0.622±0.001[c]
Aspartic acid	0.047±0.000[a]	0.270±0.000[c]	**0.291±0.001[d]**	0.201±0.000[b]
Asparagine	ND	ND	ND	ND
Arginine	0.016±0.000[a]	0.240±0.002[c]	0.224±0.001[b]	**0.260±0.000[d]**

Serine	0.017±0.000[a]	0.226±0.002[c]	**0.287±0.001[d]**	0.182±0.000[b]
Glutamic acid	0.033±0.000[a]	0.285± 0.002[b]	**0.379±0.001[d]**	0.307±0.000[c]
Glutamine	0.011±0.000[a]	0.097±0.001[c]	**0.122±0.002[d]**	0.084±0.000[b]
Glycine	0.026±0.000[a]	0.501±0.001[b]	**1.570±0.000[d]**	0.529±0.002[c]
Tyrosine	0.008±0.000[a]	0.111±0.000[b]	0.170±0.002[c]	**0.225±0.001[d]**
Cysteine	ND	ND	ND	ND
Hydroxyproline	ND	ND	ND	ND

ND- Not Detected
Values are mean of three replicates.
Means (±SD) followed the same letter are not significantly different (P≤0.05).

quality is determined by the assessment of the amino acids content and hence knowledge of the amino acid composition of foods serves as a basis for establishing their potential nutritive value [27]. Essential and non-essential amino acids ration in dietary protein has an important effect on protein utilization by fish [15]. Amino acids are associated with health issues and amino acid deficiencies lead to a number of diseases.

Essential amino acids like isoleucine, leucine, threonine and non-essential amino acids like aspartic acid, arginine, serine, glutamine and tyrosine are detected in lesser amount in all the three fish samples. Valine was detected only in *Anabas testudineus* in Pre-monsoon and Post-monsoon season. Leucine is the only dietary amino acid that can stimulate muscle protein synthesis and has important therapeutic role in stress conditions like burn, trauma and sepsis [27]. Threonine is used for treating various nervous system disorders including spinal spacticity, multiple sclerosis, familial spastic paraparesis and amyotrophic lateral sclerosis [18]. Valine is needed for the synthesis of proteins and also used as an energy fuel [24]. Arginine is essential for children's growth [2]. Aspartic acid, glutamine, proline, glycine and leucine have strong cytotoxic activity against cancer cells [35]. Tyrosine is an essential component for the production of several important brain chemicals called neurotransmitters, including epinephrine, norepinephrine and dopamine and also helps to produce melanin.

4. CONCLUSIONS

The present study shows the seasonal variations in the proximate composition and amino acids content of the three air breathing fishes *Anabas testudineus*, *Clarias batrachus* and *Channa punctata* of Loktak lake. The variation in the values might be influenced

Table 3. Amino acids profile (mg/100g) of *Channa punctata* for different seasons.

Particular	Pre-monsoon	Monsoon	Post-monsoon	Winter
Essential Amino Acids				
Histidine	0.051±0.000[b]	0.039±0.001[a]	**0.113±0.000[d]**	0.074±0.000[c]
Isoleucine	0.017±0.000[ab]	0.017±0.000[ab]	**0.019±0.002[b]**	0.016±0.001[a]
Leucine	0.051±0.000[a]	**0.108±0.002[d]**	0.100±0.001[c]	0.061±0.000[b]
Lysine	0.156±0.002[c]	0.133±0.001[b]	**0.345±0.005[d]**	0.101±0.000[a]
Metheonine	ND	ND	ND	ND
Phenylalanine	0.031±0.000[b]	0.039±0.002[c]	**0.042±0.001[d]**	0.029±0.000[a]
Threonine	0.092±0.000[a]	0.058±0.020[c]	**0.120±0.000[b]**	0.087±0.000[a]
Valine	ND	ND	ND	ND
Tryptophan	ND	ND	ND	ND
Non-essential amino acids				
Alanine	0.986±0.002[c]	0.858±0.000[b]	**1.019±0.000[d]**	0.774±0.000[a]
Aspartic acid	0.184±0.000[b]	0.178±0.001[a]	0.256±0.002[c]	**0.272±0.000[d]**
Asparagine	ND	ND	ND	ND
Arginine	0.161±0.000[b]	**0.288±0.002[d]**	0.258±0.001[c]	0.157±0.001[a]
Serine	0.168±0.002[a]	0.196±0.001[b]	**0.245±0.000[d]**	0.219±0.000[c]
Glutamic acid	0.296±0.002[b]	0.324±0.000[c]	**0.352±0.001[d]**	0.273±0.000[a]
Glutamine	0.056±0.002[b]	0.058±0.001[a]	**0.117±0.002[c]**	0.101±0.000[b]
Glycine	0.586±0.000[b]	0.484±0.002[a]	0.649±0.000[c]	**0.782±0.000[d]**
Tyrosine	**0.149±0.001[c]**	0.089±0.002[a]	ND	0.101±0.001[b]
Cysteine	ND	ND	ND	ND

Hydroxyproline	ND	ND	ND	ND

ND- Not Detected
Values are mean of three replicates.
Means (±SD) followed the same letter are not significantly different (P≤0.05).

by many factors such as seasons, species, maturity, age, sex, availability of food, environmental condition, pH, turbidity, etc. It may be concluded that these air breathing fishes are important food source of basic nutrients, protein, lipids and amino acids in all the seasons and is also able to compete with more commercially utilized species in terms of nutritional value and taste.

Moreover, further action is needed for conservation of such air breathing fishes. Unaware of their possible extinction, many habitats and feeding ground have been disturbed. Due to their unique taste, nutritive value and high demand, these air breathing fishes and their fingerlings have been sold in high price in the market. Therefore, specific government policies and research programmes are required for conservation of these air breathing fishes.

5.ACKNOWLEDGEMENT

The authors greatly acknowledge Department of Life Sciences (Zoology), Manipur University for laboratory facilities and support in carrying out this paper. SAIF, IIT Bombay for providing the instrumental facilities and supporting in amino acid analysis.

6. ABBREVIATION

L. niloticus: Labeo niloticus

B. bayad: Bagrus bayad

O. niloticus: Oreochromis niloticus

S. schall: Synodontis schall

7. REFERENCES

[1] Ackman R.G. (1989). *Nutritional composition of fats in seafood*. Progress in food & nutrition science. 13 (3-4),161-241.

[2] Adeyeye S. A. O., Fayemi O. E. and Adebayo-Oyetoro A. O. (2019). *Amino acid, vitamin and mineral profiles of smoked fish as affected by smoking methods and fish types.* Journal of Culinary Science & Technology. 17(3), 195-208.

[3] Aliyu-paiko M., Hashim R. and Amuzat A. O. (2012). *Comparison of the Whole body composition of fatty acids and amino acids between reared and wild snakehead fish Channa striata (Bloch 1793) Juveniles.* Asian Fisheries Science 25, 330-342.

[4] AOAC. (1990). *Official Methods of Analysis.* Association of Official Analytical Chemists.15th Edition. Arlington.

[5] AOAC (2000). *Official methods of Analysis*. 12th Edn. Assocoiation of Official Analytical Chemists, Washington D.C.

[6] Aremu M.O., Namo S.B., Salau R.B. Agbo C.O. and Ibrahim H. (2013). *Smoking Methods and Their Effects on Nutritional Value of African Catfish (Clarias gariepinus)*. The Open Nutraceuticals Journal. 6, 105-112.

[7] Ayanda I.O., Ekhator U.I., Bello O.A. (2019). *Determination of selected heavy metal and analysis of proximate composition in some fish species from Ogun River,* Southwestern Nigeria. Heliyon. 5 e02512.

[8] Boran G. and Karacam H. (2011). *Seasonal changes in proximate composition of some fish species from black sea*. Turk J. Fish Aquatic Soc. 11, 1-5.

[9] Chen C., Sander J.E. and Dale N.M.(2003). *The effect of dietary lysine deficiency on the immune response to Newcastle disease vaccination in Chickens.* Avians Diseases. 47(4), 1346-1351.

[10] Deka K.B., Mahanta R. and Goswami U. (2012). *Seasonal variation of protein and essential amino acid contents in Labeo gonius from Lotic and Lentic water bodies*. World J. Sci. Med. Res. 2, 71.

[11] Dogan G. and Ertan O. O. (2017). *Determination of amino acid and fatty acid composition of goldband goatfish [Upeneus moluccensis (Bleeker, 1855)] fishing from Gulf of Antalya (Turkey).* Int Aquat Res. 9, 313-327.

[12] Effiong B.N. and Mohammed I. (2008). *Effect of Seasonal variation on the nutrient composition in selected fish species in Lake Kainji Nigeria*. Nature and Science 6(2), 1-5.

[13] El Oudiani Guizani Salma and Mojahed Nizar (2015). *Atlantic Mackerel amino acids and minerals contents from the Tunisian Middle eastern coast.* International Journal of Agricultural policy and Research. 3(2), 77-83.

[14] Erkan N., Selcuk A. and Ozden O. (2010). *Amino acid and vitamin composition of raw and cooked Horse mackerel.* Food. Anal. Methods. 3, 269-275.

[15] Green J.A., Hardy R.W. and Brannon EL (2002). *The optimum dietary essential: nonessential amino acid ratio for rainbow trout (Oncorhynchus mykiss), which maximizes nitrogen retention and minimizes nitrogen excretion.* Fish. Phys. Biochem., 27, 1-2.

[16] Gulsun O. and Abdurrahman P. (2006). *Amino acid and fatty acid composition of wild sea bass (Dicentrarchus labrax): a seasonal differentiation.* European food research & technology. 222(3), 316-320.

[17] Hei A. and Sarojnalini Ch. (2012). *Proximate composition, Macro and Micro mineral elements of some smoke-dried hill stream fishes from Manipur, India.* Nature and Science. 10(1), 59-65.

[18] Hyland K. (2007). *Inherited disorders affecting dopamine and serotonin: critical neurotransmitters derived from aromatic amino acids.* Journal of Nutrition. 137(6), 1568-1572.

[19] Ishida Y., Fujita T. and Asai K. (1971). *New detection and separation method for amino acids by high-performance liquid chromatography*. Journal of chromatography. 204, 143-8.

[20] Islam M. N. and Joadder M. A. R. (2005). *Seasonal variation of the proximate composition of freshwater Gobi, Glossogobius guiris (Hamilton) from the River Padma.* Pakistan Journal of Biological Sciences. 8(4), 532-536.

[21] Jafri, A.K., D.K. Khawaja and S.Z. Qasim (1964). *Studies on the Biochemical Composition of some Freshwater fishes*.149-157.

[22] Kumaran R., Ravi V., Gunalan B., Murugan S. and Sundramanickam A. (2012). *Estimation of proximate, amino acids, fatty acids and mineral composition of mullet (Mugil cephalus) of Parangipettai, Southeast Coast of India.* Adv. Appl. Sci. Res., 3(4), 2015-2019.

[23] Lim L., Chor W. Firdaus R. F., Malitam L., Ransangan J. and Shapawi R. (2015). *Proximate and amino acid compositions of the pond-cultured spotted barb, Puntius binotatus.* Malaysian Journal of Science. 34(2), 168-171.

[24] Martin Kohlmeier (2003). *Nutrient metabolism.* Food Science and Technology, 370-377.

[25] Mohamed H. A. E., Al-Maqbaly R. and Mansour H. M. (2010). *Proximate composition, amino acid and mineral contents of five commercial Nile fishes in Sudan. African Journal of Food Science.* 4(10), 650-654.

[26] Mohammed M.O. and Alim D.I. (2012). *Amino acids contents of four commercial Nile fishes in Sudan.* African Journal of Environmental Science and Technology. 6(2), 142-145.

[27] Mohanty B., Arabinda Mahanty, Satabdi Ganguly, T.V. Sankar, kajal Chakraborty, Anandan rangasamy, Baidyanath Paul, Debajit sarma, Suseela Mathew, Kurukkan Kunnath Asha, Bijay Behera, Md. Aftabuddin, Dipesh Debnath, P. Vijayagopal, N. Sridhar, M.S. Akhtar, Neetu Sahi, Tandrima Mitra, Sudeshna banerjee, Prasenjit Paria, Debajeet Das, Pushpita Das, K.K. Vijayan, P.T. Laxmanan, and A.P. Sharma (2014). *Amino Acid composition of 27 Food Fishes and their Importance in Clinical Nutrition.* Journal of Amino Acids. 7 pages.

[28] Oluwaniyi O.O., Dosumu O.O. and Awolola GV (2010). *Effect of local processing methods (boiling, frying and roasting) on the amino acid composition of four marine fishescommonly consumed in Nigeria*. Food Chemistry.

[29] Osborne D.R. and Voogt T.P. (1978). *Analysis of nutrients in foods*. Academic Press, New York.

[30] Paul B.N., Bhowmick S., Singh P., Chanda S., Sridhar N. and Giri S.S. (2019). *Seasonal variations in proximate composition of nine freshwater fish.* Indian J. Anim. Nutr. 36(1), 65-72.

[31] Rahman M. (2021). *Estimation of Amino acids profile of Hilsa, Tenualosa ilisha in upper and lower reaches of Brahmaputra river during migration*. Journal of Global Biosciences. 10(1), 8266-8275.

[32] Rana S., Faruque Md. H., Eshik Md. M. E., Hasan Md. R. and Rahman M. S. (2019). *Seasonal variation in nutritional profile of the freshwater mud eel, Monopterus cuchia (Hamilton, 1822)*. Journal of Fisheries. 7(1), 671-680.

[33] Sarojnalini Ch. and Vishwanath W. (1988). *Nutritive value of some fishes endemic in Manipur*. Indian Journal of Fisheries. 35(2), 115-117.

[34] Saad H.A. and Alim D.I. (2014). *Amino Acids Profile of Some Economically Important Marine and Freshwater Fish From Sudan.* International Journal of Advanced Research. 3(2), 838-844.

[35] Sarma D., Das P.D., Das P., Bisht H.C.S., Akhtar M.S. and Ciji A. (2015). *Fatty acid, amino acid and mineral composition of rainbow trout (Oncorhynchus mykiss) of Indian Himalya.* Indian J. Anim. Res., 49(3), 399-404.

[36] Shearer K.D. (1994). *Factors affecting the proximate composition of cultured fishes with emphasis on salmonids.* Aquaculture, 119, 63-88.

[37] Shekhar C. (2004*). Changes in muscle biochemical composition of Labeo rohita (Ham) in relation to season.* Indian Journal of Fisheries 51(3), 319-323.

[38] SHI P. S., Wang Q., Zhu Y. T., Gu Q. H. and Xiong B. X. (2013). *Comparative study on muscle nutritional composition of juvenile bighead carp (Aristichthys nobilis) and paddlefish (Polyodon spathula) fed live feed*. Turkish Journal of Zoology. 37(3), 321-328.

[39] Singh M.B., Sarojnalini C. and Vishwanath W. (1990). *Nutritive values of sundried Esomus danricus and smoked Lepidocephalus guntea*. Food Chemistry. 36, 89-96.

[40] SPSS. (2008) *Statistical Package for Social Sciences Program*, Version 16.0 for Windows. Chicago, III, USA, SPSS Inc.

[41] Stansby, M.E. (1962). *Proximate composition of fish.* In Heen, R. and Kreuzer, R. (Eds): *Fish in Nutrition.* London Publ. News (Books) Ltd., 55.

[42] Tenyang N., Womeni H.M., Linder M., Tiencheu B., Villeneuve P. and Mbiapo F.T. (2014). *The chemical composition, fatty acid, amino acid profiles and mineral content of six fish species commercialize on the Wouri river coast in Cameroon.* La Rivista Italiana Delle Sostanze Grasse Grasse. 129-138.

[43] Yahya Bakhtiyar and Seema langer (2018). *Seasonal variation in the proximate body composition of Labeo rohita (Hamilton) from Gho-Manhasa Fish ponds.* Journal of Research & Deve;lopment. 18, 24-36.

[44] Zuraini A., Somchit M.N., Solihah M.H., Goh Y.M., Afrifah A.K., Zakaria M.S., Somchit N., Rajion M.A., Zakaria Z.A. and Mat Jais A.M. (2006). *Fatty acid and amino acid composition of three local Malaysian Channa Spp. Fish*. Food. Chem. 97, 674-678.

PREDICT THE RISK OF CARDIO VASCULAR DISEASE USING DATA MINING TECHNIQUES: A SURVEY

S.Tamil Fathima, K. Fathima Bibi[2]

[1]*Research Scholar in Computer Science, Thanthai Periyar Govt. Arts & Science College (A), Tiruchirappalli*
E-mail: tamilfathima@jmc.edu
[2]*Assistant Professor in Computer Science, Thanthai Periyar Govt. Arts & Science College (A), Tiruchirappalli*
E-mail: kfatima72@gmail.com

Abstract. *Heart disease is a very fast growing common disease and a major cause of death worldwide. Health care industry is considered as one the most information intensive industries. There is vital knowledge in the health care system and data mining technologies are commonly applied to enrich the data process. Data mining helps to make and predict the status of disease using health care data. Early detection of symptoms of heart disease is a serious challenge in the present situation. Research is being done to diagnose with hybrids of data mining techniques. The focus of this paper is to review the classification and data mining techniques which used for heart disease prediction.*

Keywords: *Heart Disease, Data Mining , Prediction, Feature Selection.*

1. INTRODUCTION

Meaningful patterns and associations of knowledgeable data can be discovered through data analysis using the process of data mining. Various methods and algorithms are applied in the data mining to extract useful knowledge in the form of pattern in the data. Knowledge Discovery Database (KDD) is considered as another name of Data mining. There are various steps involved in data mining such as Data Integration, Data Selection, Data Cleaning, Data Transformation, Data Mining, Pattern Evaluation and Knowledge Presentation and Decisions or use of Discovered Knowledge. Data mining divided into two models such as Descriptive Model and Predictive Model. Predictive Model is used to predict unknown or future values of other variables. Classification, Regression and Time-Series Analysis belong to the category of Predictive Model.

Descriptive Model is developed to issue a good knowledge of data without trying to attack a particular situation. Clustering, Association Rules, Sequence Discovery and Summarization belong to the category of Descriptive Model [1,2]. Data mining techniques such as clustering and classification methods such as Naive Bayes, Decision tree, Random forest and K-nearest neighbour are used to identify several heart diseases[3,4].

Medical Data mining is an activity of different effort which includes much inaccuracy and uncertainty [5]. Health care data is very big. The Medical Decision Support System was proposed to optimize medical errors and costs assist in earlier disease detection and to achieve preventive medicine. Some incidents will end in mistakes, huge medical cost will demage the level of support and help to the patients[6].The data mining has been used in medical domain to make medical decisions and diagnose medical problems.

2. HEART DISEASE

Heart Disease is also referred as Cardio Vascular Disease(CVD).Heart disease is one of the major risks in which the whole word is fighting on predicting the risk in the earlier stage. The human heart is one of the main components of the body, circulating blood throughout the body through the blood supply system [7].If affects the brain and heart stops working completely and life loss happens within a few minutes[8].Heart attack is caused by abrasions of the heart muscle due to insufficient oxygen supply and inattention to pump blood[9].It is difficult to accurately diagnose true heart disease because the data are so complex[10]. some of the risk factors of the Cardio vascular disease. Smoking often causes heart attack. Various factors and risk of heart disease are chest pain, heart burn and stomach pain, pain in the arms and seating. A close and updated study done in 2018 by WHO reveals the outcome that 56.9 million life loss happened in the world during the year 2016 was only by heart disease[11].The most timely tests and efficient methods for heart disease are very important.

3. FEATURE SELECTION

There are different data sets to capture heart disease, but there are some features or attributes limited no of rarely used to diagnose the disease. Some data sets have redundant features. Negative effects are caused by unwanted features. They increase the training time and so many features having redundant and irrelevant features can be very inconvenient [12,13].Feature selection is a method of removing attributes with small or no information[14].Feature selection is an efficient way to removes the unnecessary, redundant and irrelevant features from the dataset. The feature only contains relevant and useful attribute elements so it helps to reduce the training time and improve the classification accuracy[15].

4. DATA SETS

The heart disease prediction used data set taken from UCI(University of californio, Irvine) data mining repository[16]. Data set is the collection of similar data records[17]. The data set used a number of records such as1)Cleveland:303 2) Hungarian: 294 3) Switzerland: 123

4) Long Beach VA: 200. All data sets contain 76 attributes but used only 14 attributes. This 14 attributes contain eight categorical attributes and six numerical attributes. Cleveland data set and statlog data set are very popular and commonly used[18]. Because Cleveland data set and statlog data set have minimum number of missing values but other data set have more number of missing values [17].

TABLE I: Heart disease dataset Attributes used

S.NO	Attributes	Description
1	Age	Age in years
2	Sex	Gender instance 1=male, 0=female
3	cp	Chest Pain type (1= typical angina,2=atypical, 3=non- angina pain, 4=asymptomatic)
4	Trestbps	Resting blood suger (in mm Hg on admission to hospital)
5	chol	Serum cholesterol in mg/d1
6	Fbs	Fasting blood sugar>120 mg/d1(1=true, 0=false)
7	restecg	Resting electrocardiographic results(0= normal, 1= having ST-T wave abnormality, 2=left ventricular hypertrophy)
8	thalach	Maximum heart rate
9	exang	Exercise Induced Angina (1=yes, 0=no)
10	oldpeak	ST Depression include by Exercise Relative to Rest
11	slope	Slope of the Peak Exercise ST Segment(1 = Up Sloping 2 = Flat 3 = Down Sloping)
12	ca	Number of Major Vessels Colored by Fluoroscopy
13	thal	Defect types 3 = Normal 6 = Fixed Defect 7 = Reversible Defect
14	num	Diagnosis of heart disease Class (0 = Healthy 1 = Have Heart Disease)

5. DATA MINING TECHNIQUES USED IN HEART DISEASE PREDICTION

Karthikeyan G et. al [19] proposed a hybridization method which combines linear stacking model and Xgboost algorithm (HLS-Xgboost) to improve prediction accuracy. This HLS-Xgboost model performs better than other models such as Decision Tree (DT), Naive Bayes (NB) classifier, Density base Spatial clustering model, SVM, Random Forest (RF), Multi-Layer perceptron (MLP), and Linear Regression (LR) respectively. It gives 96% more of accuracy than other existing models.

Vicky Singh et. al [20] proposed a recommendation system using machine learning algorithms based on vital data like cholesterol level, age, etc. for heart disease prediction. The proposed model gives 90% of accuracy and performs better than SVC and Decision tree classifier.

Md. Nahiduzzaman et. al [21] proposed two classifiers such as Multi-Layer Perceptron neural network (MLP) and another is Support Vector Machine (SVM). In addition, the heart disease is classified into two-class and five-class level in the proposed work. SVM, MLP gives 92.45%, 90.57% accuracy respectively in two-class classification problem. SVM, MLP produces accuracy of 59.01%, 68.86% in five-class classification problem. They concluded that SVM outperforms in two-class classification and MLP Performs better in five-class classification for heart disease diagnosis.

Devansh Shah et. al [22] presented a model which depends on supervised learning algorithms. K- nearest neighbor algorithm gives 90.7% of more accurate result than other supervised learning algorithms such as Naïve Bayes, decision tree, random forest.

Saba Bashir et. al [23] described the heart disease prediction using feature selection techniques and algorithms to enhance accuracy. Logistic regression SVM, Decision Tree, Naïve Bayes and Random forest are used in Rapid miner tool on a UCI dataset. The experimental result shows that the accuracy of Decision Tree, Regression, Random Forest, Naïve Bayes and Logistic Regression SVM of 82.22%, 82.56%, 84.17%, 84.24% and 84.85% respectively. They suggested
that Logistic Regression (SVM) is a better feature selection technique for heart disease prediction.

.Kanika Pahwa et. al [24] proposed a heart disease prediction model using hybrid feature selection SVM-RFE (support Vector Machine - Recursive Feature Elimination) algorithm for feature selection. UCI dataset been used in the proposed model. In addition, Naive Bayes, Random Forest were used as classifiers for categorize the disease existence or absence.

C. Sowmiya et. al [25] proposed a hybrid model for heart disease prediction using ACO (Ant Colony Optimization) with HKNN (Hybrid K-Nearest Neighbor) classifier. This ACO-HKNN is compared with existing KNN (K-Nearest Neighbor), C4.5, Naïve Bayes, Decision Tree and Support Vector Machine (SVM) classification techniques. The proposed hybrid prediction model produced a better accuracy of 99.2% when compared with other existing classification techniques.

Pooja Rani et. al [26] proposed a hybrid verdict carry system by combining the GA (Genetic Algorithm) and recursive feature elimination for feature selection. Synthetic Minority Oversampling Technique and standard scalar technique used for pre-processing. SVM, naive bayes, random forest, logistic regression and adaboost classifiers are used in the proposed system. It provides 86.6%, of accuracy.

Anchana Khemphila et. al [27] proposed an improved approach for Heart disease Classification using MLP (Multi Layer Perceptron) with BPLA (Back-Propagation Learning

Algorithm) and Feature Selection Algorithm (FLA). Attributes were reduced from Thirteen to eight and the accuracy differences is 1.1% in training data set and 0.82% in the validation data set.

Namariq Ayad Saeed et. al [28] proposed a Heart Disease Prediction System by combining BPSO (Binary Particle Swarm Optimization Algorithm) with Mutual Information (MI) filter. Logistic regression is one of the most important technique used for classification. This MI_BPSO produces a Classification accuracy of 98.33% when compared with BPSO. Execution time of the MI_PBSO been significantly improved when compared with existing BPSO.

Luxmi Verma et. al [29] proposed a hybrid model using CFS (Correlation based Feature Subset) with PSO (Particle Swam Optimization) and Kmeans clustering algorithms for diagnosis of Coronary artery disease (CAD). The proposed hybrid model provides 90.82% accuracy than other existing techniques such as MLR (Multinomial Logistic Regression), FURIA (Fuzzy Unordered Rule Induction Algorithm), MLP (Multi-Layer Perceptron) and C4.5.

Youness Khourdifi et. al [30] proposed the FCBF (Fast Correlation-Based Feature) selection to enhance the heart disease classification worth. In addition, the existing classification algoritms such as KNN, SVM, Multilayer Perception, Artificial Neural Network, NB, RF are optimized by PSO (Particle Swarm Optimization) combined with ACO (Ant Colony Optimization). This hybrid optimized model gives the 99.65% of classification accuracy.

J. Vijayashree et. al [31] presented a novel function based on population diversity and tuning function for identifying optimal weights. In addition, they proposed a fitness function for PSO with SVM in order to reduce attributes count and the accuracy improvement. The proposed PSO- SVM produces better accuracy than other existing feature selection algorithms.lications domain, strength and weakness of their algorithms.

M. Anbarasi et. al [32] proposed the GA with three classifiers like NB, classification by clustering and DT used to predict the diagnosis of patients with reduced number of features. Thirteen attributes are reduced to six attributes by using GA. The NB and classification by clustering having inconsistencies and high missing value but DT data mining techniques accu-racy is high 99.2% and less missing value.

Durgadevi Velusamy et. al [33] proposed a machine learning algorithm for effective diagnosis and prediction of CAD (Co-ronary Artery Disease). It contains a heterogeneous ensemble method which combines the classifiers such as Random For-est, KNN and SVM for effective diagnosis and ensemble vot-ing technique for CAD prediction. In feature selection, the Boruta wrapper based feature selection algorithm and SVM have been used based on attribute importance and rank. In ensemble voting technique, among based Majority-Voting (MVEn), Average Voting (AVEn), and Weighted-Average Voting (WAVEn), the WAVEn algorithm gives 98.97% of classification accuracy, 100% of sensitivity, 96.3% of speci-ficity, and 98.3% of precision for the original dataset. In the balanced dataset, the WAVEn algorithm achieves 100% of accuracy, sensitivity, precision and specificity in CAD diagnosis.

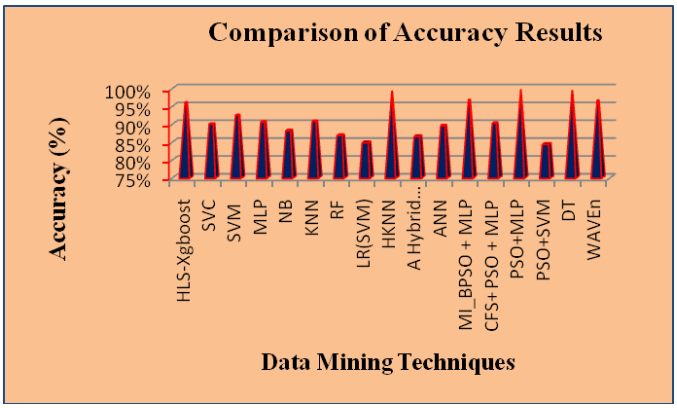
Jyoti Soni et. al [34] are focused on a detailed survey related to heart disease prediction using data mining techniques. In same dataset, Decision Tree, Bayesian Classification algorithm are performed better than other prediction models such as Neural Networks, KNN, Classification based on clustering. In addition, the accuracy of Bayesian classification, Decision Tree algoritm have been further improved by combining Genetic Algorithm. The comparison of all the techniques is summarized as given in Table 2.

Table 2: The accuracy accrued by the existing data mining techniques used for the prediction of CVD was summarized in the given below.

Author	Data Mining Techniques	Accuracy	Feature Selection
Karthikeyan et. al	HLS-Xgboost	96%	13
Vicky Singh et. al	SVC	90%	13
Md.Nahiduzzaman et. al	SVM MLP	92.45% 90.57%	13
Devansh et. al	NB KNN DT RF	88.157% 90.789% 80.263% 86.84%	13
Saba Bashir et. al	MRMR/FS LR(SVM)	84.85%	13
Kanika Ravinderkumar et. al	NB RF	84.1584% 84.164%	10 12
Sowmiya.C et. al	HKNN Hybrid KNN	99.02%	9
Pooja Rani et. al	A Hybrid combining GA& RFE	86.60%	8
Anchana Khemphila et. al	ANN	Training 89.56%	8
Namariq Ayad saeed et. al	MI_BPSO + MLP	96.66%	8
Luxmi Verma et. al	CFS+ PSO + MLP	90.28%	7
Youness Khourdifi et. al	PSO+MLP	99.65%	7
Vijayashree et. al	PSO+SVM	84.36%	6
M.Anbarasi et. al	DT	99.2%	6
Durgadevi Velusamy et. al	WAVEn	96.55%	5

6. RESULTS AND DISCUSSIONS

The comparative analysis among the existing data mining techniques HLS-Xgboost, SVC, SVM, MLP, NB, KNN, DT, RF, LR(SVM), HKNN, A Hybrid combining GA& RFE, ANN, MI_BPSO + MLP, CFS+ PSO + MLP, PSO+MLP, PSO+SVM, WAVEn is carried out. The performance of the existing data mining techniques are compared with respect to the accuracy results is as shown in the Fig.1.



7. CONCLUSION

Knowing the initial stage of the heart disease with data mining techniques is one of the biggest challenges in the health care sector. In health care sector different processes generate data in huge quantities. Some of the computerized health care monitoring systems and various medical instruments are continuously collecting health care data and hence the volume of the clinical data is increasing in rapid manner. Early detection of symptoms of heart disease can save people from this disease. The important task of this paper is to review the classification and feature selection of data mining techniques used for heart disease prediction. The techniques used in reviews produced good outcomes. This will be beneficial to the patients with heart disease if the accuracy is achieved by hybriding the available techniques.

8. REFERENCES

[1] Venkatadri.M. Dr.Lokanathan C.Reddy (2011), A Review on Data Mining from Past to Future, International journal of Computer Applications, Vol 15(7), pp.19-22.

[2] Leventhal, Barry (2010), An introduction to data mining and other techniques for advanced analytics. Journal of Direct, Data and Digital Marketing Practice 12, no. 2 (2010): 137-153.

[3] Rafiah Awang and Palaniappan.S (2007), Web based Heart Disease Decision Support System Using Data Mining Classification Modeling Techniques, Proeeding of iiWAS, pp.177-187.

[4] Patel J, TejalUpadhyay D, Patel S (2015), Heart disease prediction using machine learning and data mining technique. Heart Disease. Vol 7(1),pp.129–37.

[5] Monali Dey,Siddharth Swarup Rautaray (2014), Study and Analysis of Data Mining Algorithms for Healthcare Desicion Support System, International Journal of Computer Science and Information Technology, Vol 5(1), pp.470-477.

[6] Siri Krishnan Wasan, Vasutha Bhatnagar and Harleen Kaur(2006), The Impact of Data Mining techniques on medical diagnostics, Data Science Journal, Vol 5(19), pp 119-126.

[7] Tanya Lewis. (2016, March 22). Human Heart: Anatomy, Function & Facts [Online]. Available: https://www.livescience.com/34655 human-heart.html.

[8] Keerthana, T.K(2017),Heart disease prediction system using data mining method,Int. J. Eng. Trends Technol. Vol 47(6), pp 361–363.

[9] Sudha.A, Gayathri.P and Jaishankar.N(2012), Utilization of Data MiningvApproaches for prediction of life Threatening Disease Survivability, IJAC(0975-8887).

[10] Alia, A.F, Tawee (2017), A Feature selection based on hybrid binary cuckoo search and rough set theory in classification for nominal datasets. Inf. Technol. Comput. Sci. 4(April), 63–72.

[11] World Health Organization. The world health report 2000: health systems: improving performance. World Health Organization, 2000.

[12] Mohammad Ashraf Ottom, Girija Chetty, Dat Tran, and Dharmendra Sharma(2012), Hybrid approach for diagnosing thyroid, hepatitis, and breast cancer based on correlation based
feature selection and nave bayes. volume 7666, pages 272– 280.

[13] Dharmendra Modha and W. Spangler(2003). Feature weighting in k-means clustering. Machine Learning, Vol 52:217–237.

[14] Latha Parthiban and R.Subramanian(2007), Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm, International journal of Biological and Life Sciences, Vol 3(3), pp.157-160.

[15] TaoWang, ZhenxingQin,ZhiJin and ShichaoZhang(2010). Handling over-fitting in test cost-sensitive decision tree learning by feature selection, smoothing and pruning. Journal of Systems and Software, 83:1137–1147.

[16] Kurgan, Lukasz A., Cios, Krzysztof J., Tadeusiewicz, Ryszard, Ogiela, Marek, & Doodenday, Lucy S. (2001), Knowledge discovery approach to automated cardiac SPECT diagnosis. Artificial Intelligence in Medicine, 149-169.

[17] "UCI Machine Learning repository:", http://archive.ics.uci.edu/ml/datasets/ Heart Disease.
[18] Israa Nadheer1, Mohammad Ayache2, Hussein Kanaan(2021), Heart Disease Prediction System Using Machine Learning Algorithm, Journal of Advanced Computer Science & Technology, Volume 8, pp.23-31.
[19] Karthikeyan G, Komarasamy G, Daniel Madan Raja S(2021), An Efficient Method for Heart Disease Prediction Using Hybrid Classifier Model in Machine Learning, Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pp. 5708 – 5717.
[20] Vicky Singh, Brijesh Pandey(2021), Prediction of Cardiac Arrest and Recommending Lifestyle Changes to Prevent It using Machine Learning, International Conference on Intelligent Technologies & Science, pp. 1-6.
[21] Md. Nahiduzzaman, Md. Julker Nayeem, Md. Toukir Ahmed(2019), Prediction of Heart Disease Using Multi-Layer Perceptron Neural Network and Support Vector Machine, 4th International Conference on Electrical Information and Communication Technology (EICT), Khulna, pp 1-6.
[22] Devansh Shah, Samir Patel, Santosh Kumar Bharti(2020), Heart Disease Prediction using Machine Learning Techniques, SN Computer Science, pp. 1-6.
[23] Saba Bashir, Zain Sikander Khan, Farhan Hassan Khan, Aitzaz Anjum, Khurram Bashir(2019),Improving Heart Disease Prediction Using Feature Selection Approaches, Proceedings of 2019 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, pp. 619-623.
[24] Kanika Pahwa, Ravinder Kumar(2017), Prediction of Heart Disease Using Hybrid Technique For Selecting Features, 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON) GLA University, Mathura, pp. 500-504.
[25] C. Sowmiya, P. Sumitra(2020), "A hybrid approach for mortality prediction for heart patients using ACO- HKNN", Journal of Ambient Intelligence and Humanized Computing, pp. 1-8.
[26] Pooja Rani, Rajneesh Kumar, Nada M. O. Sid Ahmed, Anurag Jain(2021), A decision support system for heart disease prediction based upon machine learning, Journal of Reliable
Intelligent Environments, Springer, pp. 1-13.
[27] Anchana Khemphila, Veera Boonjing(2011), Heart disease Classification using Neural Network and Feature Selection, 21st International Conference on Systems Engineering, IEEE Computer Society, pp. 406-409.
[28] Anchana Khemphila, Veera Boonjing(2011), Heart disease Classification using Neural Network and Feature Selection, 21st International Conference on Systems Engineering, IEEE Computer Society, pp. 406-409.
[29] Luxmi Verma, Sangeet Srivastava(2016),P. C. Negi, A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data, J Med Syst, vol 40, pp. 1-7.
[30] Youness Khourdifi, Mohamed Bahaj(2019), Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization, International Journal of Intelligent Engineering and Systems, Vol.12, No.1, 2019, pp. 242-252,
http://www.inass.org/10.22266/ijies2019.0228.24
[31] J. Vijayashree, H. Parveen Sultana(2018), A Machine Learning Framework for Feature Selection in Heart Disease Classification Using Improved Particle Swarm Optimization with Support Vector Machine Classifier, Programming and Computer Software, 2018, Vol. 44, No. 6, pp. 388–397, https://doi.org/10.1134/S0361768818060129
[32] M. Anbarasi , N.ch.s.n.iyengar ,n.ch.s.n.iyengar(2010), Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm , International Journal of Engineering Science and Technology Vol. 2(10), ISSN: 0975-5462, pp. 5370-5376.
[33] Durgadevi Velusamy, Karthikeyan Ramasamy(2021), Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset, Computer Methods and Programs in Biomedicine, Elsevier, Vol. 198, 2021, pp. 1-13, https://doi.org/10.1016/j.cmpb.2020.105770
[34] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni(2011), Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, International Journal of Computer Applications (0975 – 8887), Vol. 17– No.8, pp. 43-48.