ISSN: 2355-6579 | E-ISSN: 2528-2247
DOI: https://doi.org/10.31294/inf.v10i1.12787

Avoiding Machine Learning Becoming Pseudoscience in Biomedical Research

Meredita Susanty[1], Ira Puspasari[2], Nilam Fitriah[3], Dimitri Mahayana[4], Tati Erawati Latifah Rajab[5], Hasballah Zakaria[6], Agung Wahyu Setiawan[7], Rukman Hertadi[8]

[1] Universitas Pertamina
Jl Teuku Nyak Arief Simprug Kebayoran Lama DKI Jakarta, Indonesia

[2] Universitas Dinamika
Jl. Raya Kedung Baruk No.98, Kedung Baruk Rungkut Surabaya Jawa Timur, Indonesia

[3,4,5,6,7,8] Institut Teknologi Bandung
Jl. Ganesa No.10, Lb. Siliwangi Coblong, Bandung Jawa Barat, Indonesia

e-mail: [1]meredita.susanty@universitaspertamina.ac.id, [2]33221050@std.stei.itb.ac.id, [3]33220307@std.stei.itb.ac.id, [4]dimitrimahayanastei@gmail.com, [5]tati@stei.itb.ac.id, [6]fahala@gmail.com, [7]awsetiawan@itb.ac.id, [8]rukman@chem.itb.ac.id

Article Information Submitted: 20-04-2022 Revised: 07-10-2022 Accepted: 15-02-2023

Abstract

The use of machine learning harbours the promise of more accurate, unbiased future predictions than human beings on their own can ever be capable of. However, because existing data sets are always utilized, these calculations are extrapolations of the past and serve to reproduce prejudices embedded in the data. In turn, machine learning prediction result raises ethical and moral dilemmas. As mirrors of society, algorithms show the status quo, reinforce errors, and are subject to targeted influences – for good and the bad. This phenomenon makes machine learning viewed as pseudoscience. Besides the limitations, injustices, and oracle-like nature of these technologies, there are also questions about the nature of the opportunities and possibilities they offer. This article aims to discuss whether machine learning in biomedical research falls into pseudoscience based on Popper and Kuhn's perspective and four theories of truth using three study cases. The discussion result explains several conditions that must be fulfilled so that machine learning in biomedical does not fall into pseudoscience.

Keywords: deep learning; philosophy; biomedical

1. Introduction

Machine learning is the study of algorithms that improve their performance at some tasks from experience (Mitchell, 1997). While traditional programming uses data and programs to produce an output, machine learning uses data and output to produce a program. The main goal of a learner is to generalize their experiences. In this context, generalization refers to a learning machine's ability to accurately execute new, previously unseen examples/tasks after observing a learning data set (Bishop, 2006). Observation of existing data is performed iteratively to generate a predictive model.

Allowing the computer to "decide" what is relevant within the parameters specified eliminates many detrimental human biases and allows less space for researcher assumptions about an association or cause-and-effect relationship in the generation of a model. The training examples are drawn from an unknown probability distribution. The learner must develop a general model of this space that will allow it to make sufficiently accurate predictions in new cases. Because training sets are lim-ited and the future is uncertain, learning theory rarely guarantees algorithm performance. Probabilistic performance bounds are pretty common. The decomposition of bias and variance is one technique to measure generalization error.

While machine learning shows less human bias, other sorts of biases emerge. Due to the model's large capacity, machine learning algorithms are capable of forming unrealistic relationships among variables. When the algorithm memorizes the training data due to these unrealistic connections, it is known as overfitting. Overfitting might be caused by relying on limited measurements and failing to validate the data correctly. Furthermore, such algorithms are data hungry, and supplying small amounts of data can easily lead to model overfitting. Machine learning fairness, model generalizability, and model drifting are among the other drawbacks.

In biomedical fields, image recognition, object detection, 3D reconstruction, and other medical image processing are computer vision problems that can be solved by machine learning (Park et al., 2018). Machine learning is classified into supervised learning (e.g. classification) and unsupervised learning (e.g. clustering). Nowadays, biomedical researchers applying digital image processing to extract, analyze, and classify Magnetic Resonance Imaging (MRI) results comprehensively review tumours. The research for brain tumor type discrimination using MRI features has developed parameters of brain tumor classification (Iqbal et al., 2018). The other application is diabetic retinopathy classification (Mansour, 2018) using a convolutional neural network (CNN) with 97.93% accuracy. The other biomedical image analysis assists the doctors in polyp detection (Billah & Waheed, 2018), which showed that CNN features and colour wavelets could highly represent endoscopic polyp images with an accuracy of 98.23%.

Besides biomedical image processing, machine learning also has been widely used in biomedical signal processing or protein structure study (bioinformatics). Automatic heart activity diagnosis based on Gram Polynomials and Probabilistic Neural Networks was developed (Beritelli et al., 2018). The Deep Neural Network (DNN) was also implemented to detect REM, where the data is collected by one channel electrocardiography (ECG) (Wei et al., 2017). Bolland et al. implemented DeepBind with CNN to predict specificities of DNA- and RNA-binding proteins (Bolland et al., 2016).

Since machine learning results highly depends on the data fed in into the model, some are likely to be inaccurate or wrong because the software is identifying patterns that exist only in that data set and not the real world. Machine learning and statistical techniques that shift through large amounts of data make uncertain in their results and unlikely reproducible. The application of an algorithm that compounds existing inequities in socioeconomic status, race, ethnic background, religion, gender, disability, or sexual orientation and amplifies inequities in health systems (Chen et al., 2008; Larrazabal et al., 2020; Ledford, 2019; Lyratzopoulos et al., 2013; Madabhushi & Lee, 2016; Maya Dusenbery, 2018; Meghani et al., 2012; Panch, Mattie, & Atun, 2019; Panch, Mattie, & Celi, 2019; Pelletier et al., 2014) can be considered as a pseudoscience. Instead of revealing the truth in the real world, the machine learning result confirms the social inequities in society as truth.

Telling real science apart from pseudoscience is not easy. Pseudoscience may look good on the surface, but dig a bit deeper, and its fake claims are simply too good to be true. This article focused on the avoidance of machine learning application in biomedical research to become pseudoscience. We discussed the demarcation from the perspective of Karl Popper's view and Thomas Kuhn's. Although both views have different arguments about science, they have the essential contribution in differentiating science and non-science (or pseudoscience). The discussion would cover the science's definition from Karl Popper and Thomas Kuhn and the theory of truth as the fundamental purpose of the research. Moreover, we included the three cases in biomedical research to support the understanding and enrich the discussion.

2. Research Method

2.1. Karl Popper’s Criteria of Science

Pseudoscience is something that resembles a science while it is not a science indeed. For example, an astrologist shows that they can identify the biographical information of great people from their zodiac signs (FS, 2021). Most world leaders and celebrities are Leos, which astrologically tend to be ambitious, strong, and attention seekers. With adequate and supportive sample observations, it looks like a scientific theory.

Karl Popper (1902-1994), a philosopher of science, felt that calling Marxist theory and psychoanalysis a science is not proper, in contrast to viewing Newton theory or Einstein theory of relativity as a science. He gave an example regarding his doubts (Popper, 1963) on Freud's psychoanalytic theory and its relation to Alfred Adler's theory. There are two men to be compared. The first man pushes a kid to make him drown. The second man gets into the water to save the kid who is drowning. According to Freud, the first person suffers from pressure while the second one experiences an increase in behavior level. Besides, according to Adler's theory, the first person felt inferiority that he wanted to have the courage to do something bad while the second person also proved himself brave enough to save the drowning child. Popper argued that human behaviour could not be interpreted with either theory.

From Popper's formulation of science (Popper, 1963), the psychoanalytic theory from Freud and Adler is not testable because there is no human behaviour that can refute it. This psychological theory may be of considerable importance, but not in a testable form yet. The same thing happens to astrology. Meanwhile, Einstein's Theory of Relativity regarding gravity meets the criteria of testability. Even with limited equipment in his time, there is a possibility to be refuted/tested. Nowadays, physicists still try deriving it to achieve more understanding, and that theory is still testable and accurate. Thus, Popper argues that science should be testable means it is also risky of being proven false. Popper (Popper, 1963) stated the definition of testability as follows: If observation shows that the predicted effect is absent, then the theory is easily refuted.

Popper also points out that testability is not a problem how meaningful a theory is or how acceptable it is. However, Popper makes testability the dividing line between science and non-science (including pseudoscience), as illustrated in Fig.1 (a). He called the testability criteria a solution to the demarcation problem.

2.2 Thomas Kuhn's Scientific Revolution

Thomas Kuhn argued that science does not evolve gradually towards the truth that established theories were simply overturned and replaced with new ones. Science has a paradigm that remains constant before going through a paradigm shift when existing theories fail to explain a phenomenon and someone suggests a new theory.

According to Kuhn, the progression of science is not linear but alternates between 'normal' and revolutionary' (or 'exceptional') phases. Revolutionary periods are not simply periods of rapid advancement; they are essentially different from normal science. On the surface, normal science resembles the conventional cumulative picture of scientific progress. Kuhn describes normal science as "puzzle-solving." The puzzle-solver expects to have a reasonable chance of succeeding, that his success will be based primarily on his abilities, which means it involves subjectivity, and that the puzzle and its methods of solution will have a high degree of familiarity. Normal science can expect to accumulate a growing stock of puzzle solutions. On the other hand, scientific revolutions are not cumulative since they involve a revision of current scientific theory or practice (G. E. Jones, 1981). In a revolution, not all of the achievements of the previous period of normal science are preserved. A later period of science may find itself without explaining a phenomenon that was thought to be satisfactorily explained in a previous period. This characteristic of scientific revolutions is known as 'Kuhn loss.'

Thomas Kuhn mapped the stages of the development of science into four main phases (Thomas S . Kuhn, 1990), pre-paradigm phase, normal science phase, crisis phase, and scientific revolution phase. In the pre-paradigm phase, as the immature science phase, scientific study on specific topics is carried out here without any defined goals or objectives in mind. Various types of thoughts emerge throughout this period, competing with and excluding one another. To be considered a science, a scientific field must attain a consensus in the shade of a particular paradigm. Kuhn believed that once an agreement was formed, scientists began to engage in the normal science phase. A commitment to establishing a shared paradigm that will determine the rules of the game and all standard benchmarks in scientific practice is a normal science precondition. Outside of the current paradigm, "normal" scientists will not make any discoveries. Instead, they are actively engaged in applying the paradigm to understand natural signs in greater depth better. The moment when knowledge can no longer be depended upon to solve problems as they arise is defined as the crisis phase or the phase of the emergence of extraordinary sciences. The scientific community began to question the dominant paradigm. During a crisis, one of the emerging ideas will be able to overcome scientific challenges, generalize, and promise a future of improved scientific study. At this time, extraordinary sciences are no longer extraordinary. This transition is the scientific revolution phase, a non-cumulative developmental experience in which a new one partially or entirely replaces an older paradigm.

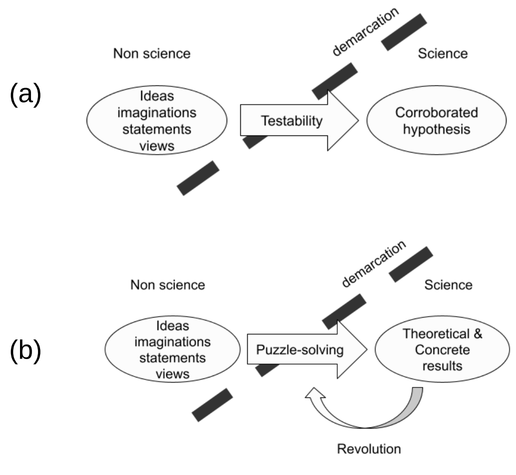

Fig 1. (a) Popper's and (b) Kuhn's demarcation

2.3. The Theories of Truth

The correspondence theory of truth (David, 2018) — that whatever corresponds to observable reality is true. If there is an adequate fact to which a belief corresponds, it is true, and vice versa. With facts and structured propositions in hand, an attempt may be made to explain the relation of correspondence. The most obvious application of this theory is in science: it is used in an experiment to refute a hypothesis because it is assumed that what is observed in the experiment is what is true.

The coherence theory of truth (Walker, 2018) — the claims are true if they follow logically and coherently from a set of axioms (or intermediate propositions). Arguments must make sense which they must flow logically from premises and intermediate propositions. The truth conditions of propositions, according to the coherence theory, are built from other propositions. Meanwhile, according to the correspondence theory, the truth conditions are objective features of the world rather than propositions.

The consensus theory of truth (Bufacchi, 2021; Hesse, 1978)— that what is true is what everyone agrees to be true. This idea is flawed for the following reasons: even if everyone reaches an agreement, all of them may be wrong. It can only be reached through an idealized discourse process even though the consensus is a crucial component of the scientific method. We are frequently unable to verify the correctness of a scientific study independently. Instead, we put our faith in science's system to reach a provisional conclusion on what is true and known about the world.

The pragmatic theory of truth (Capps, 2020)— that what is true is what is helpful to you. The "practical repercussions" are more important than theoretical ones, and it is the epistemology of the practitioner. The pragmatic theory also has some intriguing implications for argument evaluation. Suppose the standard criterion of truth is usability, and there are sophisticated arguments sound right but do not work. In that case, a logical next step is to filter arguments based on the trustworthiness of the person saying them.

It is crucial to highlight that none of these truth theories is superior to the others, and they are all acceptable for some forms of truth but not for others; they are all flawed. The objective of explaining the four theories is to demonstrate that humans apply different standards of truth to different situations.

3. Result and Discussion

With the perspective of Karl Popper and Thomas Kuhn about science, we elaborated the risk of becoming pseudoscience in biomedical research in three cases and discussed how to keep this research to be categorized as science, not pseudoscience. First, we identified the abnormality of human health that is classifying abnormal heart sound to assist clinicians in decision making. Second, we used the task in the rehabilitation process assisted by the computer, named human computer interaction (HCI). Specifically, we took brain computer interface (BCI), which became one of the active studies in biomedical research. Last, a bioinformatics task to predict protein structure.

3.1 Normal and abnormal heart sound classification

Nowadays, the number of people with heart disease has increased globally, including 13.6% of the population in China, 22% in Canada, 26.3% in Egypt, and 50 million in the United States. In Indonesia, 6-15% of the population suffers from heart disease (Jon Christian, 2019). In addition, countries in North America and Europe have more than 80% of heart failure patients above 65 years old. It is estimated that in 2030, heart failure patients in the United States will reach 8.5 million (Erickson, 2003).

The heart is a vital organ, which has functions to pump blood continuously throughout the body. It consists of the myocardium muscle, i.e. two atria and two ventricles. Typical heart sound for adults consists of two signals, the first sound (S1) is usually associated with cardiac vibrations due to the closure of mitral and tricuspid valves, and the second sound (S2) is a result of cardiac vibrations produced by the closure of the aortic and pulmonic valves. Other sounds produced by the heart due to structural and functional defects are called murmurs(Erickson, 2003). Phono-cardiography is a non-invasive technique that records heartbeat patterns, including heart sounds and murmurs. The record is in a time-domain series of heart sound signals as a phonocardiogram (PCG). Heart sound consists of two phases: systolic and diastolic phases in cardiac cycles. Fig.2 shows signals for normal (above), and abnormal heart sounds signals (below). As can be seen in the top part of Fig.2(a), the most fundamental heart sounds are the first and the second (S1 and S2, respectively) sounds.

On the other hand, a PCG of a type of abnormal heart sound signal is depicted in Fig.2(b). It shows the two fundamental components of the signals perturbed by murmur. In this figure, called mitral regurgitation murmur. Murmur, being a pseudo-periodic non-stationary high-frequency signal, needs both temporal and spatial trends to discriminate the disease from normal classification. It will also reduce biases such as ambience and motion artefacts which are relatively aperiodic.

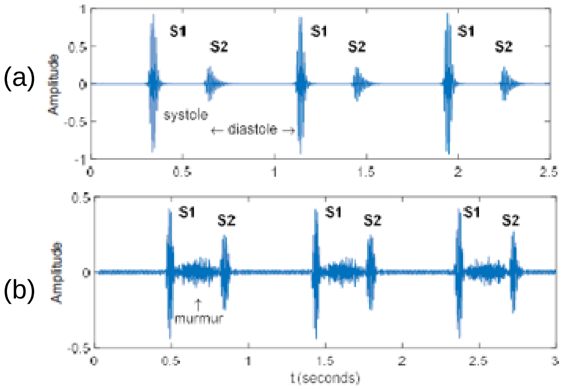

Fig 2. (a) Normal and (b) Abnormal Heart Sound Signal [31]

Identification of heart sound defects requires deep analysis using various signal processing methods: the acquisition of heart sound signals, noise elimination (denoising), feature extraction, and classification. Artificial intelligence has helped in several things in research, especially in biomedicine, especially for PCG signal analysis. Implementing feature extraction using the Ensemble method and integration with deep learning using 124 time-frequency features resulted in classification for abnormal signal recognition by accuracy level 0.8602, providing recommendations for the screening process (Puspasari et al., 2019). Implementing the RNN method, LSTM, GRU, B-RNN, B-LSTM, and CNN by utilizing three layers of Deep learning on the Physionet Dataset produces 80% accuracy (Potes et al., 2016). Firuzbakht et al. (Firuzbakht et al., 2018) proposed the SVM method to detect ab-normal heart sound, consisting of 23 subjects with five normal subjects, six subjects Aortic Stenosis, 12 Tricuspid Regurgitation, obtained an accuracy of 96.2%. It becomes pseudoscience if immeasurable subjectivity is applied in its implementation.

Classification of PCG heart signals, normal and abnormal classification develops ANN by the integration of various models. There are classification stages: filtering and segmentation. Based on correspondence theory, this research is incompatible because each process cannot be touched even sensed—so many PCG samples and data sets, impossible for researchers to calculate manually. One process that follows the correspondence truth theory is when the heart sound signal retrieval directly uses a stethoscope instead of a database. This theory requires that the five senses can capture each stage of research. Signal processing is based on the heart sound signal, which is indeed the signal that can be seen, processed but cannot be touched. For example, characteristic extraction based on the frequency value cannot be seen since it computerized processing.

Further research refers to the consistency of the previous studies. Pragmatic truth theory does not attach importance to correlation and correspondence truth, as long as a proposition has benefits. What if one method of classification is better than the others and provides optimal results? During the study, several methods of characteristic extraction of the heart sound signal in the time and frequency domain showed the characteristics of normal heart sound, including S1 and S2 components and sytolic and diastolic murmurs. Model algorithm of heart sounds abnormality identification was built by training machine learning with weights from previously trained CNN and extensive data to achieve higher accuracy.

Kuhn argued that science goes through the scientific revolution. Similarly, various machine learning applications will experience a revolution to achieve either a high level of accuracy or a low level of computing.

3.2 Brain signals pattern recognition (specific case in speech production)

Human living activity produces unique signals. One of them is bioelectric signals, besides biomagnetic signals, bioacoustic signals. Bioelectric signals are generated by the nerve cells and muscle cells, while the cell membrane generates an action potential under certain conditions (Onaral & Cohen, 2006). This signal can be acquired by electrodes placed on the surface or invasively near the cell. One tool that has been developed is electroencephalography (EEG) to acquire bioelectric signals from the brain. EEG with surface electrodes has the lowest recorded potential due to the resistance of the scalp and skull. Despite its limitation, from the early 1300s, EEG signals have carried the development of clinical studies, experiments, and computational works for detection, recognition, diagnosis, and physical treatment of many neurological problems and physiological abnormalities of the brain.

One field that also uses EEG contribution is neurocognitive study, for example, how humans feel fully concentrated or very painful, or how humans use a language to speak, as illustrated in Fig.3. Machine learning has been employed in most neurocognitive studies to identify the EEG pattern. This kind of study contains a pragmatic truth since the primary purpose is its usability. Additionally, the neurocognitive study is often used to help people with disabilities to perform their activities in the form of the Brain Computer Interface (BCI), such as to control a wheelchair, type a letter, or

Internet of Things (IoT) applications. While people can observe the actual use of the result, this study contains the correspondence truth.

Related to the importance of determining whether machine learning application in the neurocognitive study is a science, this part specifically discussed the intention of EEG pattern recognition with machine learning to fall into the pseudoscience. Based on Popper, testability is required for science. In the neurocognitive study, pattern recognition must observe samples that represent nearly the variety of data. Indeed, variety in bioelectric signals is not always about labels of some specific activities. They also vary due to the users (cross-subject or inter-subject) and the personal condition (intra-subject). For example, several imagined speech recognitions based on brain signals could gain accuracy of above 90% in specific persons (Parhi & Tewfik, 2021; Saha & Fels, 2019), conducted with the same dataset (Nguyen et al., 2018). Even if the accuracy is high, the data was gathered from 15 subjects at one time, so there was no time-variety consideration.

Additionally, there was a lack of cross-subject analysis, and this high accuracy potentially tends to be overfitting. Nowadays, some researchers try to adopt transfer learning to answer the problem of cross-subject (Cooney et al., 2019; García-Salinas et al., 2019). Although, the accuracy is still low.

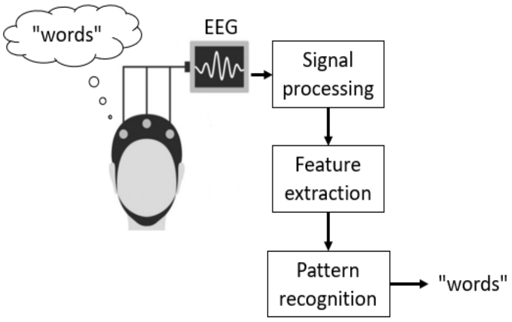

Fig 3. Imagined speech recognition based on brain signals

The other main issue for neurocognitive study with machine learning is how well the model could explain the phenomenon of brain activity. Although the accuracy is high enough after we employed massive parameter tuning, we have no guarantee that the model could achieve the same level in other uncovered varieties. For imagined speech recognition based on the brain signals, the variety could be due to the language used, other equipment besides EEG, e.g., magnetoencephalography (MEG, which acquired brain biomagnetic signals) or microelectrode (invasively placed), or brain condition within the subjects. While the phenomenon is discovered, the causality of different results for the new domains can be well observed.

Kuhn has a different perspective about science demarcation from Popper, where Kuhn stated that the capability of puzzle-solving is the fundamental property of normal science. Thus, in Kuhn's point of view, this is not about testability but problem solving capability. Kuhn also has stated that science is revolutionary, just like machine learning in pattern recognition in the neurocognitive study that also goes through revolution. Most models in pattern recognition are often refined as many problems arise.

In terms of speech pattern recognition, in the early phase, research has been conducted to recognize the pattern of an acoustic signal produced by a human. There are still problems to be answered, such as noise cancellation, language limitation. However, several models have been found to work well in speech technology, such as Google Assistant. If we revisit the history of speech pattern recognition, it started from observing how people speak and recognizing the information from the speech (meaning, dialect, gender). Then, it becomes the normal science while it struggles to answer the problem. Thus, it needs the scientific revolution.

Pattern recognition in speech has entered the crisis period to find the solutions for the anomalies. One of the anomalies is how if a person cannot produce a voice when he speaks? For example, the one who undergoes a tracheostomy, a person with articulator abnormality, or a paralyzed one who relies only upon his brain activity (Bocquelet et al., 2016). The biomedical study has evolved that observation of speech has been beyond the acoustic signal; there are muscular signals in the articulator and brain signals as the source (Schultz et al., 2017). In this case, the neurocognitive study is needed to observe the speech production from the source (brain activity).

Pattern recognition of speech from acoustic and brain signals is different since their frequency range is different; brain signal in 0.1-100Hz (Onaral & Cohen, 2006) and acoustic signals in 100-10,000Hz (Ryan & Frater, 2002). Additionally, the brain signal of speech production is analogous as an order before it was translated to muscular activity to produce the sound of speech. The acoustic signal is affected by the air produced by the lung, length of the vocal tract and its mass, condition of the vocal fold, and the activity in the mouth tract and nasal tract (Anusuya & Katti, 2011). In contrast, brain signals represent the cognitive activity when humans think to speak involving the network of millions of neurons (Chang et al., 2015). They are affected by human focus and the acquisition limitation of electrodes.

These days, pattern recognition in brain signals, including in the speech production process, has been actively implemented to answer the problems related to human beings. The scientific revolution aimed to solve the puzzle and enrich the existing technology, e.g. speech technology (Krishna et al., 2019).

3.3 Protein Structure Prediction

Protein structure prediction (PSP) has long been a significant issue in biochemistry. Protein structures are critical for understanding the fundamental biology of health and disease and becomes the basis for other studies such as drug development. Computational approach is used as an alternative to determining protein structure when experimental techniques such as X-ray crystallography (Slabinski et al., 2007), NMR spectroscopy (Markwick et al., 2008), and increasingly, cryo-electron microscopy (Jonic & Vénien-Bryan, 2009) are limited. It does not replace the experimental one. In this case, computational approaches are rooted in the pragmatic theory of truth, which is frequently concerned with instrumental outcomes; it is less concerned with the specifics of why something works.

Modern PSP systems typically consist of four components (AlQuraishi, 2021):(i) an input module that takes a single protein sequence and generates additional input features, such as multiple sequence alignment (MSA) of homologous proteins (Senior et al., 2019, 2020), Myogenic regulatory factors (MRF) (Golkov et al., n.d.; Liu et al., 2018), pairwise potential (D. T. Jones & Kandathil, 2018; Wang et al., 2017), Position Specific Scoring Matrix (PSSM) (Alquraishi, n.d.; Ingraham et al., 2018; Xu et al., 2020), (ii) a machine learning model for recognize pattern, that transforms features from the input to spatial information that partially encodes the 3D structure using various architecture such as Convolutional (Golkov et al., n.d.; D. T. Jones et al., 2015), Residual Net (ResNet) (Senior et al., 2020; Wang et al., 2017), Attention mechanism (Jumper et al., 2021) (iii) an output that converts this spatial information into a Binary contact map (D. T. Jones et al., 2015; Wang et al., 2017), Distogram (Senior et al., 2020; Xu, 2018), Orientogram (Yang et al., 2020), or preliminary 3D (AlQuraishi, 2018; Ingraham et al., 2018; Jumper et al., 2021) structure, sometimes without explicit side-chain atoms, and (iv) a refinement module that improves the preliminary structure and generates all atomic coordinates (Jumper et al., 2021). PSP is undergoing a paradigm shift. These modules have traditionally relied on a combination of physics-based energy functions, knowledge-based statistical reasoning, and heuristic algorithms. However, there has been an infusion of machine learning, particularly neural networks, into every aspect of PSP in recent years. According to Kuhn's phase, traditional modules, which were the normal science, were unable to predict the protein structure with acceptable accuracy. PSP has been in the crisis phase since CASP1 in 1994 because no model can provide a promising result. In 2018, a model (AlphaFold) using ResNet architecture (Senior et al., 2020) outperformed all other models in CASP13. It signalled the start of the revolution phase. In 2020, another model (AlphaFold2) using a Transformer based architecture (Jumper et al., 2021) improved the former model and achieved micro-angstrom accuracy. Since then, machine learning in PSP has become the normal science. Many scientists use Transformer architecture to predict RNA structure, which has shown significant success in predicting protein structure. Although AlphaFold2 is a game-changer, it can only predict a single protein structure. Many other use cases are still being explored or are in the crisis phase. These are multi-chain prediction, disordered or unstructured region prediction, the effect of mutation prediction, multiple conformations in protein folding, and positions of any non-protein components found in experimental protein structure.

Protein structure prediction is defined as a well-defined problem with precise inputs and outputs: predict the 3D structure (output) given AA sequences (input), with experimental structures serving as the ground truth (labels). The proposition that protein structure from experimental results is the ground truth for the computational approach follows the coherence theory of truth. The protein structures from experimental technologies conform with various laws, theories, and dogma in biology, physics, and chemistry. Thus, it is logical that the prediction from the computational approach must align with the ones from the experimental approach. Protein sequences, like human language, can be naturally represented as sequence of word. The protein sentence is made up of 20 standard AAs considered as words. Furthermore, like natural language, naturally evolved proteins are typically composed of reused modular elements with minor variations that can be rearranged and assembled hierarchically. The completeness of the information is another crucial feature shared by proteins and human language. A sequence of AAs determines its three-dimensional structure and function. It means that the protein's information (e.g., structure) is contained within its sequence, according to information theory. Because of these similarities in shape and substance, many PSP uses natural language processing (NLP) methods to predict protein structure from sequence (Ofer et al., 2021). Although the current PSP system using deep learning achieves near-angstrom accuracy for single protein prediction, neural networks are frequently referred to as "black boxes". The highly recursive structure makes the resulting parameters and functions too complex for practitioners to understand. Even though we know the amino acid sequences of billions of proteins and their final three-dimensional structure, predicting how they get it is challenging.

In contrast to machine learning/deep learning in social science inferences, which may become pseudoscience if not handled carefully, machine learning/deep learning in PSP is less likely to become pseudoscience. First, most of the scientists in this area use verified and standardized data from publicly available databases. Second, it takes the protein structure from the experimental results as a ground truth. Third, most scientists make their model's code publicly available so that others can evaluate and verify their findings. Fourth, scientists hold a biannual event called Critical Assessment of Protein Structure Prediction to track the progress of PSP (CASP). The protein structure used in CASP is a structure that has been measured experimentally but has not been published so that the competition participants do not know the 3D structure of the protein (4). The prediction result is assessed using the Global Distance Test (GDT) (Zemla, 2003) to assess prediction accuracy. GDT provides accurate measurement than the typical root means square deviation (RMSD) metric – which is sensitive to outlier regions created, for example, by poor modelling of individual loop regions in a structure that is otherwise reasonably accurate. The GDT score is calculated as the largest set of amino acid residues' alpha carbon atoms in the model structure falling within a defined distance cut-off of their position in the experimental structure after iteratively superimposing the two structures. These efforts ensure that the model is falsifiable and testable. Criteria that must be met so that a scientific achievement can be considered as science.

4. Conclusion

It is undoubtedly that machine learning plays a role in identifying the pattern that is hard to see with the naked eyes and needs years of experience. Moreover, it also avoids human errors in tasks. As machine learning has been used widely in the biomedical field, one mistake could have a huge risk for human life. For example, when a clinician decides the patient's heartbeat pattern looks normal while it is not, then it makes the patient's condition worse; when a surgeon makes a wrong diagnosis by reading the CT scan result, this could lead to wrong operation action; a mistake in analyzing cognitive activity in human being causes the wrong understanding then loses its benefits; a fault in human's protein structure prediction will bring other errors in the usage. Due to the cruciality, we have to make sure that machine learning is applied scientifically.

Based on Popper's criteria of science, science has to be testable intended to falsify it, not to support it. Thus, the theory becomes corroborated. In this term, a biomedical scientist has to be aware of the testing procedure aimed for falsification. This work could refer to any machine learning problem, such as overfitting, fairness, and data drifts since machine learning depend on the data. For example, we should feed the machine learning model with high variability data related to the task. By publishing the model and dataset built, we invite other researchers to falsify it. Besides, the prior knowledge of experts would help to build the explainable model. Additionally, as time goes by, we cannot ignore that the data trends could shift, causing the model are no longer applicable. Adaptivity of the model is the other important thing.

While Popper gets rid of subjectivity in science demarcation by requiring testability, Kuhn argues that science still needs subjectivity. In contrast, science is revolutionary; then this involves a consensus to accept it. As we mentioned about fairness and expert involvement, it means biomedical research matures the part of the scientific revolution to answer the "puzzles," and biomedical expert's involvement fulfils the subjectivity aspect in science acceptance. The aim of publications is also to open the chance of more scientific revolutions.

We can conclude that applying machine learning in biomedical research is not as easy as crunching the data. Many things need consideration due to enhance the quality of human life. We cannot ignore the role of consensus while machine learning in biomedical science goes through the scientific revolution to answer biomedical questions. Thus, we also have to be firmed that biomedical research is testable and aware of any machine learning challenge.

References

Alquraishi, M. (n.d.). *End-to-end Differentiable*

Learning of Protein Structure. https://doi.org/10.1101/265231

AlQuraishi, M. (2018). End-to-end differentiable learning of protein structure. In *bioRxiv* (p. 265231). bioRxiv. https://doi.org/10.1101/265231

AlQuraishi, M. (2021). Machine learning in protein structure prediction. *Current Opinion in Chemical Biology*, *65*, 1–8. https://doi.org/10.1016/j.cbpa.2021.04.005

Anusuya, M. A., & Katti, S. K. (2011). Front end analysis of speech recognition: A review. In *International Journal of Speech Technology* (Vol. 14, Issue 2). https://doi.org/10.1007/s10772-010-9088-7

Beritelli, F., Capizzi, G., Lo Sciuto, G., Napoli, C., & Scaglione, F. (2018). Automatic heart activity diagnosis based on Gram polynomials and probabilistic neural networks. *Biomedical Engineering Letters*, *8*(1), 85. https://doi.org/10.1007/S13534-017-0046-Z

Billah, M., & Waheed, S. (2018). Gastrointestinal polyp detection in endoscopic images using an improved feature extraction method. *Biomedical Engineering Letters*, *8*(1), 69. https://doi.org/10.1007/S13534-017-0048-X

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics): .: : Amazon.com: Books*. Springer. https://www.amazon.com/Pattern-Recognition-Learning-Information-Statistics/dp/0387310738

Bocquelet, F., Hueber, T., Girin, L., Chabardès, S., & Yvert, B. (2016). Key considerations in designing a speech brain-computer interface. *Journal of Physiology-Paris*, *110*(4), 392–401. https://doi.org/10.1016/j.jphysparis.2017.07.002

Bolland, D. J., Koohy, H., Wood, A. L., Matheson, L. S., Krueger, F., Stubbington, M. J. T., Baizan-Edge, A., Chovanec, P., Stubbs, B. A., Tabbada, K., Andrews, S. R., Spivakov, M., & Corcoran, A. E. (2016). Two Mutually Exclusive Local Chromatin States Drive Efficient V(D)J Recombination. *Cell Reports*, *15*(11), 2475–2487. https://doi.org/10.1016/J.CELREP.2016.05.020

Chang, E. F., Raygor, K. P., & Berger, M. S. (2015). Contemporary model of language organization: an overview for neurosurgeons. *Journal of Neurosurgery*, *122*(2), 250–261. https://doi.org/10.3171/2014.10.JNS132647

Chen, E. H., Shofer, F. S., Dean, A. J., Hollander, J. E., Baxt, W. G., Robey, J. L., Sease, K. L., & Mills, A. M. (2008). Gender disparity in analgesic treatment of emergency department patients with acute abdominal pain. *Academic Emergency Medicine : Official Journal of the Society for Academic Emergency Medicine*, *15*(5), 414–418. https://doi.org/10.1111/J.1553-2712.2008.00100.X

Cooney, C., Folli, R., & Coyle, D. (2019). Optimizing Layers Improves CNN Generalization and Transfer Learning for Imagined Speech Decoding from EEG. *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 1311–1316. https://doi.org/10.1109/SMC.2019.8914246

Erickson, B. (2003). *Heart sounds and murmurs across the lifespan* (4th ed.). Mosby.

Firuzbakht, F., Fallah, A., Rashidi, S., & Khoshnood, E. R. (2018). Abnormal Heart Sound Diagnosis Based on Phonocardiogram Signal Processing. *26th Iranian Conference on Electrical Engineering, ICEE 2018*, 1450–1455. https://doi.org/10.1109/ICEE.2018.8472410

FS. (2021). Karl Popper on The Line Between Science and Pseudoscience. In *Farnam Street Media Inc*. https://fs.blog/2016/01/karl-popper-on-science-pseudoscience/

García-Salinas, J. S., Villaseñor-Pineda, L., Reyes-García, C. A., & Torres-García, A. A. (2019). Transfer learning in imagined speech EEG-based BCIs. *Biomedical Signal Processing and Control*, *50*, 151–157. https://doi.org/10.1016/j.bspc.2019.01.006

Golkov, V., Skwark, M. J., Golkov, A., Dosovitskiy, A., Brox, T., Meiler, J., & Cremers, D. (n.d.). *Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images*. Retrieved November 8, 2021, from http://hmmer.org

Ingraham, J., Riesselman, A., Sander, C., Marks, D., & School, H. M. (2018). *Learning Protein Structure With A Differentiable Simulator*.

Iqbal, S., Khan, M. U. G., Saba, T., & Rehman, A. (2018). Computer-assisted brain tumor type discrimination using magnetic resonance imaging features. *Biomedical Engineering Letters*, *8*(1), 5. https://doi.org/10.1007/S13534-017-0050-3

Jon Christian. (2019). *Statistician: Machine Learning Is Causing A "Crisis in Science."* https://futurism.com/neoscope/machine-learning-crisis-science

Jones, D. T., & Kandathil, S. M. (2018). High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics*, *34*(19), 3308–3315. https://doi.org/10.1093/bioinformatics/bty341

Jones, D. T., Singh, T., Kosciolek, T., & Tetchner, S. (2015). MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics (Oxford, England)*, *31*(7), 999–1006. https://doi.org/10.1093/BIOINFORMATICS/BTU791

Jones, G. E. (1981). Kuhn, Popper, and Theory Comparison. *Dialectica*, *35*(4), 389–397. https://doi.org/10.1111/j.1746-8361.1981.tb00791.x

Jonic, S., & Vénien-Bryan, C. (2009). Protein structure determination by electron cryo-microscopy. *Current Opinion in Pharmacology*, *9*(5), 636–642. https://doi.org/10.1016/J.COPH.2009.04.006

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*. https://doi.org/10.1038/s41586-021-03819-2

Krishna, G., Tran, C., Yu, J., & Tewfik, A. H. (2019). Speech Recognition with No Speech or with Noisy Speech. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1090–1094. https://doi.org/10.1109/ICASSP.2019.8683453

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, *117*(23).

Ledford, H. (2019). Millions of black people affected by racial bias in health-care algorithms. *Nature*, *574*(7780), 608–609. https://doi.org/10.1038/D41586-019-03228-6

Liu, Y., Palmedo, P., Ye, Q., Berger, B., & Peng, J. (2018). Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. *Cell Systems*, *6*(1), 65-74.e3. https://doi.org/10.1016/j.cels.2017.11.014

Lyratzopoulos, G., Abel, G. A., McPhail, S., Neal, R. D., & Rubin, G. P. (2013). Gender inequalities in the promptness of diagnosis of bladder and renal cancer after symptomatic presentation: evidence from secondary analysis of an English primary care audit survey. *BMJ Open*, *3*(6), e002861. https://doi.org/10.1136/BMJOPEN-2013-002861

Madabhushi, A., & Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, *33*, 170–175. https://doi.org/10.1016/J.MEDIA.2016.06.037

Mansour, R. F. (2018). Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy. *Biomedical Engineering Letters*, *8*(1), 41. https://doi.org/10.1007/S13534-017-0047-Y

Markwick, P. R. L., Malliavin, T., & Nilges, M. (2008). Structural Biology by NMR: Structure, Dynamics, and Interactions. *PLOS Computational Biology*, *4*(9), e1000168. https://doi.org/10.1371/JOURNAL.PCBI.1000168

Maya Dusenbery. (2018, May 29). *"Everybody was telling me there was nothing wrong" - BBC Future*. https://www.bbc.com/future/article/20180523-how-gender-bias-affects-your-healthcare

Meghani, S. H., Byun, E., & Gallagher, R. M. (2012). Time to Take Stock: A Meta-Analysis and Systematic Review of Analgesic Treatment Disparities for Pain in the United States. *Pain Medicine*, *13*(2), 150–174. https://doi.org/10.1111/J.1526-4637.2011.01310.X/2/PME_1310_F8IK.JPEG

Mitchell, T. M. (1997). *Machine Learning* (1st ed.). McGraw-Hill Education.

Nguyen, C. H., Karavas, G. K., & Artemiadis, P. (2018). Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features. *Journal of Neural Engineering*, *15*(1), 16002. https://doi.org/10.1088/1741-2552/aa8235

Ofer, D., Brandes, N., & Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. In *Computational and Structural Biotechnology Journal* (Vol. 19, pp. 1750–1758). Elsevier B.V. https://doi.org/10.1016/j.csbj.2021.03.022

Onaral, B., & Cohen, A. (2006). Biomedical Signals. In J. D. Bronzino (Ed.), *Medical Devices and Systems* (3rd ed., pp. 1–22). CRC Press. https://doi.org/10.1201/9781420003864.sec1

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems. *Journal of Global Health*, *9*(2). https://doi.org/10.7189/JOGH.09.020318

Panch, T., Mattie, H., & Celi, L. A. (2019). The “inconvenient truth” about AI in healthcare. *Npj Digital Medicine 2019 2:1*, *2*(1), 1–3. https://doi.org/10.1038/s41746-019-0155-4

Parhi, M., & Tewfik, A. H. (2021). Classifying imaginary vowels from frontal lobe EEG via deep learning. *European Signal Processing Conference*, *2021-January*, 1195–1199. https://doi.org/10.23919/EUSIPCO47968.2020.9287599

Park, C., Took, C. C., & Seong, J. K. (2018). Machine learning in biomedical engineering. *Biomedical Engineering Letters 2018 8:1*, *8*(1), 1–3. https://doi.org/10.1007/S13534-018-0058-3

Pelletier, R., Humphries, K. H., Shimony, A., Bacon, S. L., Lavoie, K. L., Rabi, D., Karp, I., Avgil Tsadok, M., & Pilote, L. (2014). Sex-related differences in access to care among patients with premature acute coronary syndrome. *CMAJ*, *186*(7), 497–504. https://doi.org/10.1503/CMAJ.131450/-/DC1

Popper, K. R. (1963). *Conjectures and Refutations*. Routledge and Keagan Paul. https://eportfolios.macaulay.cuny.edu/liu10/files/2010/08/KPopper_Falsification.pdf

Potes, C., Parvaneh, S., Rahman, A., & Conroy, B. (2016). Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. *Computing in Cardiology Conference*, 621–624. https://ieeexplore.ieee.org/document/7868819

Puspasari, I., Kusumawati, W. I., Oktarina, E. S., & Jusak, J. (2019). A New Heart Sound Signal Identification Approach Suitable for Smart Healthcare Systems. *Proceedings of the 2019 2nd International Conference on Applied Engineering, ICAE 2019*. https://doi.org/10.1109/ICAE47758.2019.9221752

Ryan, M. J., & Frater, M. (2002). *Communications and Information Systems*. Argos Press.

Saha, P., & Fels, S. (2019). Hierarchical Deep Feature Learning for Decoding Imagined Speech from EEG. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 10019–10020. https://doi.org/10.1609/aaai.v33i01.330110019

Schultz, T., Wand, M., Hueber, T., Krusienski, D. J., Herff, C., & Brumberg, J. S. (2017). Biosignal-Based Spoken Communication: A Survey. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *25*(12), 2257–2271. https://doi.org/10.1109/TASLP.2017.2752365

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2019). Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins: Structure, Function and Bioinformatics*, *87*(12), 1141–1148. https://doi.org/10.1002/prot.25834

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

Slabinski, L., Jaroszewski, L., Rodrigues, A. P. C., Rychlewski, L., Wilson, I. A., Lesley, S.

A., & Godzik, A. (2007). The challenge of protein structure determination--lessons from structural genomics. *Protein Science : A Publication of the Protein Society*, *16*(11), 2472–2482. https://doi.org/10.1110/PS.073037907

Thomas S . Kuhn. (1990). The Road since Structure. *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *2*, 3–13.

Wang, S., Sun, S., Li, Z., Zhang, R., & Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Computational Biology*, *13*(1), e1005324. https://doi.org/10.1371/journal.pcbi.1005324

Wei, R., Zhang, X., Wang, J., & Dang, X. (2017). The research of sleep staging based on single-lead electrocardiogram and deep neural network. *Biomedical Engineering Letters*, *8*(1), 87–93. https://doi.org/10.1007/S13534-017-0044-1

Xu, J. (2018). Distance-based Protein Folding Powered by Deep Learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(34), 16856–16865. http://arxiv.org/abs/1811.03481

Xu, J., McPartlon, M., & Li, J. (2020). Improved protein structure prediction by deep learning irrespective of co-evolution information. In *bioRxiv*. bioRxiv. https://doi.org/10.1101/2020.10.12.336859

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., & Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(3), 1496–1503. https://doi.org/10.1073/PNAS.1914677117/-/DCSUPPLEMENTAL

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Research*, *31*(13), 3370–3374. https://doi.org/10.1093/nar/gkg571