	いすらむ たんじら
氏　　名	**ISLAM TANZILA**
本籍（国籍）	バングラデシュ人民共和国
学位の種類	博士(工学)
学位記番号	理工博　第１２号
学位授与年月日	令和５年３月２３日
学位授与の要件	学位規則第５条第１項該当　課程博士
研究科及び専攻	理工学研究科システム創成工学専攻
学位論文　題目	**A Deep Learning Method to Impute Missing Values and Compress Genome-wide Polymorphisms for Predicting Phenotype of Rice** **(イネの表現型予測に向けた欠測値推定とゲノムワイド多型圧縮のための深層学習法)**
学位審査委員	主査　准教授　　木村　彰男 副査　教授　　西山　　清 副査　教授　　永田　仁史 副査　東京大学大学院　准教授　　岩田　洋佳

論　文　内　容　の　要　旨

Genomic selection (GS) is expected to accelerate plant and animal breeding, but during the last decade, genome-wide polymorphism data have increased enormously, which has raised concerns about storage cost and computational time. This necessitates the development of innovative platforms that can considerably reduce the resources required for storage and processing. Several individual studies, such as imputing missing genome data, compressing genome data, and predicting phenotypes, have attempted to solve those concerns. However, there still have remained some issues: imputation models are developed only for imputing missing genotypes, compression models lack adequate quality of data after compression, and prediction models are time-consuming. Therefore, this thesis proposes a novel combined deep learning model that could resolve these limitations and outperform the other state-of-the-art methods. The proposed model is roughly composed of three parts: (i) AGI (autoencoder genome imputation) for imputing missing genome-wide polymorphisms, (ii) AGC (autoencoder genome compression) for compressing genome-wide polymorphism data by generating different level of compression according to storage

requirements, and (iii) DeepCGP (deep learning compression-based genomic prediction) for predicting phenotypes from compressed information using random forests, genomic best linear unbiased prediction, and Bayesian variable selection.

The organization of this thesis is as follows.

Chapter 1 introduces the background of the research, declares the research motivation and objectives, reviews related works and the structure of the dissertation. In particular, it is stated that there have been no prediction methods that can predict the phenotypes of a target trait based on compressed genome-wide polymorphism data in animal and plant breeding.

Chapter 2 discusses imputing missing genome-wide polymorphism data. Missing data consists in the lack of information in a dataset that directly influences data analysis performance. Hence in this thesis a deep learning model, AGI has been proposed which can impute missing genome-wide polymorphism data using deep autoencoder. By applying pre-processing technique to the original data, the dataset is split into several parts and trained the separated autoencoder network to reduce the computational time. The AGI model is evaluated by comparing other existing imputation methods such as Simple Imputer and BEAGLE, and the result shows AGI model can achieve up to 96% accuracy to impute missing genome-wide polymorphisms from a multiparent advanced generation intercross (MAGIC) population dataset.

Chapter 3 considers genome-wide DNA polymorphism data compression. In particular, AGC, a deep learning model, has been proposed to compress genome-wide polymorphism data using a separated deep autoencoder. The aim is to compress high-dimensional data and retain high quality information from the compressed data that can be used for any kind of data analysis instead of original data. The compression levels of AGC can be adjusted depending on the storage requirements, which is substantial for resources and computational time. The experiment shows AGC model can compress Cornell-IR LD rice array data (C7AIR), a one of famous dataset for rice breeding and genetics studies, up to 94.01% and another famous high-density rice array dataset (HDRA) up to 98.57%.

Chapter 4 proposes the DeepCGP to predict the phenotype of rice. The DeepCGP consists of two models: the first one is an autoencoder model for compressing genome-wide polymorphism data, and the second one is a regression model that takes the compressed information generated by the autoencoder as input and attempts to predict the genotypic values of a target trait. Existing regression models such as random forests (RF), genomic best

linear unbiased prediction (GBLUP), and Bayesian variable selection (BayesB) can also be used for prediction. Experiments present DeepCGP can obtain up to 96% prediction accuracy after 94.01% compression for C7AIR and 99% prediction accuracy after 98.57% compression even for high-density HDRA, which demonstrates the high quantitative quality of the AGC compression method, too. Among three regression models BayesB shows the highest accuracy; however, it requires extensive computational time for high-dimensional original uncompressed data such as HDRA and could only be used with compressed data.

In Chapter 5, the performance applying convolutional neural network (CNN) with DeepCGP for predicting phenotypes is discussed. From evaluations after carefully selecting the network parameters to minimize the effect of overfitting, it is demonstrated that CNN-based learning can potentially improve the accuracy over traditional prediction methods and can assist in predicting phenotypes.

Chapter 6 presents the comparative analysis of predictive performance of RF, BayesB, GBLUP and CNN. By comparing the predictive performance for the original uncompressed data and different levels of compressed data for two datasets C7AIR and HDRA, and overall, CNN shows better performance both in compressed data and original uncompressed data among the four models for genomic prediction.

Chapter 7 concludes the thesis and presents recommendations for future research.

論 文 審 査 結 果 の 要 旨

世界人口の増加が続く中，持続的かつ安定的な食料増産は人類の喫緊の課題であり，増産には作物の育種が重要である。一般的な育種は，交配や選抜を繰り返しながら栽培試験を行うため，10年程度の長い期間を要するとされているのに対し，本研究では，大幅な育種期間の短縮を期待できるゲノム情報を用いた手法に着目している。これは，ゲノムデータが栽培初期の苗の段階でも抽出できることによる。一方，ここ数十年の間に明らかにされたゲノムデータは巨大かつ複雑になっており，データの分析／保存／転送のための時間的／ストレージ的コストが無視できない，という問題も同時に抱えていた。このため，ゲノム解析に必要な計算機リソースを大幅に削減できるような，革新的なプラットフォームが望まれている。

このような背景の下，本論文では，ゲノム解析に有用な複合型の深層学習モデルを新たに提案している。提案モデルは，(i) ゲノムワイド多型データの欠損を推定して穴埋めするためのAGI (Autoencoder Genome Imputation)，(ii) ゲノムワイド

多型データをストレージ要件に応じて異なる圧縮レベルで圧縮するための AGC (Autoencoder Genome Compression),および（iii）圧縮されたゲノムワイド多型データから，育種に関連する（収量や品質等の）重要形質の表現型を高い精度で予測するための DeepCGP (Deep learning Compression-based Genomic Prediction)，の 3 つから構成されている。

第 1 章では，研究背景，研究動機，研究の目的などが述べられており，特に，動植物育種において，圧縮されたゲノムワイド多型データから対象形質の表現型を予測する方法がこれまでに存在していなかったことが述べられている。

第 2 章では，ゲノムワイド多型の欠損データのインピュテーション（遺伝型の推定）について述べられている。既存のゲノムデータセットはしばしば情報が欠落しており，その欠損がゲノムデータ解析に悪影響を与えてしまうことから，本論文では，これに対処するための方法として，ディープオートエンコーダに基づく深層学習モデル AGI を提案し，これによって欠損データを推測している。AGI は，元のデータセットに前処理を施して複数に分割し，それらを個別に学習させることで計算規模の縮小と処理時間の短縮を図っている。この AGI モデルは，MAGIC データセット（多親子世代間交配集団のゲノムワイド多型データ）に対し，欠損の予測を最大 96%の精度で実現できることを実験的に実証している。その性能は，Simple Imputer や BEAGLE 等の既存の予測法を凌いでおり，高く評価できる。

第 3 章では，ゲノムワイド多型のデータ圧縮法として，分離型のディープオートエンコーダによって可変的にデータを圧縮できる深層学習モデル AGC を提案している。AGC の特長は，高次元のデータを効率よく圧縮できることに加え，圧縮後のデータにおいても，さまざまなデータ解析での利用に耐えうる高品質な情報を保持できることにある。評価実験においては，イネの品種改良や遺伝学研究で著名な Cornell-IR LD Rice Array データ（C7AIR）を 94.01%（元データの約 6%程度まで），同じく高密度イネ配列データ（HDRA）を 98.57%（元データの約 1.5%程度まで）圧縮できることを示しており，これらは高く評価できる。

第 4 章では，圧縮されたゲノムワイド多型データからイネの表現型を予測するための深層学習モデル DeepCGP を提案している。DeepCGP の入力は AGC が生成した圧縮情報，出力は対象形質の遺伝子型（予測値）であるが，予測のために用いる具体的な回帰手段として，ランダムフォレスト（RF），ゲノム最良線形予測（GBLUP），ベイズ選択（BayesB）といった既存モデルが利用できるように設計している。そして，この DeepCGP の予測精度は，C7AIR データにおいて 94%圧縮後に最大で 96%，HDRA データにおいて 98%圧縮後に最大で 99%であることを評価実験によって示している。つまり，元データの僅か数%程度のデータ量から極めて高精度な予測が実現でき，AGC 圧縮によるデータ品質の低下がほぼないことを定量的に示したといえる。このような予測は DeepCGP によって初めて実現された，といっても過言ではなく，これらの成果は極めて高く評価できる。

第 5 章では，DeepCGP をベースとした畳み込みニューラルネットワーク（CNN）を

新たに構築しており，これをイネの表現型予測に適用すると高い性能を示すことを評価実験によって実証している。さらに第 6 章では，DeepCGP の予測モデルとして，RF，GBLUP，BayesB，CNN をそれぞれ用いた場合の性能について，比較が試みられている。具体的には，C7AIR と HDRA の 2 つのデータセットについて，元の非圧縮データと異なるレベルの圧縮データ 3 種に対する予測精度と処理速度を比較しており，圧縮データ／非圧縮データのいずれにおいても，CNN が全体的に良い性能を示すことを述べている。このように，CNN の導入によって DeepCGP の性能をさらに改善させたことは高く評価できる。

第 7 章では，結論とともに，今後の課題や展望が述べられている。

以上，本論文は，ゲノムワイド多型データの欠損推定，データ量の大幅圧縮，および圧縮データからの重要形質の表現型予測を可能とする，複合型の深層学習モデルを新たに提案したものであり，従来は実現されていなかった圧縮ゲノムデータからの表現型予測法を初めて確立させたという点で，知能情報工学分野の発展に寄与するところが大きい。よって，本論文は博士（工学）の学位論文として合格と認める。

原著論文名（１編を記載）

DeepCGP: A Deep Learning Method to Compress Genome-wide Polymorphisms for Predicting Phenotype of Rice
Tanzila Islam, Chyon Hae Kim, Hiroyoshi Iwata, Shimono Hiroyuki and Akio Kimura
IEEE/ACM Transactions on Computational Biology and Bioinformatics (掲載決定済)
DOI: 10.1109/TCBB.2022.3231466