Machine Learning Models for
Predicting Potato Variety Performance Across Environments

Md Munimul Islam | C00317189@setu.ie | MSc – Applied Artificial Intelligence | SETU Carlow
Supervisor: **Libor Zachoval** | libor.zachoval@setu.ie

INTRODUCTION

The potato (Solanum tuberosum L.) is a critical global food crop facing mounting pressure from climate change and growing demand for food security.

Teagasc, Ireland's Agriculture and Food Development Authority, has conducted a multi-location potato breeding programme since 1962, successfully developing 30 varieties (e.g. Rooster). This programme generates extensive multi-location trial (MET) datasets capturing yield, quality, and disease responses under diverse environmental conditions.

The Challenge:

Current MET workflows rely on traditional linear statistical models (AMMI, GGE biplots) which:

- Treat environment as a categorical variable
- Cannot capture complex nonlinear genotype × environment (GEI) interactions
- Limit early variety selection effectiveness

Machine learning and deep learning methods offer the potential to model these nonlinear interactions — but lack the explainability required for trustworthiness and adoption.

RESEARCH QUESTION

"Can explainable machine learning models, trained on Teagasc historical multi-location trial data, improve the prediction of potato variety performance (yield – kg/ha) and stability across environments, compared to traditional statistical linear models?"

Sub-Questions:

- Which ML/DL models best balance accuracy, LOEO stability & explainability?
- Can XAI frameworks (SHAP) identify key GEI drivers for variety performance vs. statistical models?
- What are the benefits and limitations within Teagasc's potato breeding context?

DATA SOURCE

Teagasc Genovix Database

Historical multi-location potato trial data collected since 2004 across multiple sites in Ireland & globally.

Key Feature Categories: (Total 257 Features)

- Trial Identifiers — Pedigree, Location, Year, Plot, Block, Cross ID
- Yield & size grading (44) — FINALYIELD, YIELDTHA
- Tuber Quality traits (13) — Dry Matter, Tuber Size, Tuber Shape, Uniformity, Eye Depth, Appearance
- Disease & defect (40) — BLIGHT, DRYROT, PVY, BLACKLEG
- Storage traits (9 cols) — STOR_APPR, STOR_FIRM, STOR_FFSPR, STOR_DORM
- Cooking & processing (22) — Crisp, Chip Color, Mealiness, Flavour

Environmental covariates (per site/year) : weather summaries (temperature, humidity, rainfall), soil proxies (pH,) – WILL BE AUGMENTED

STATE OF THE ART

Statistical Methods | **Machine Learning** | **Deep Learning**

Statistical Models (Baseline):

AMMI & GGE biplots are standard in Teagasc's workflow. Effective for variance partitioning and linear GEI, but treat environment as categorical and cannot model nonlinear trait relationships (Piekutowska & Niedbała, 2025).

Machine Learning:

Ensemble models (Random Forest, XGBoost, LightGBM) consistently outperform linear baselines. Yu et al. (2025) achieved MAPE < 10% using an explainable ensemble with SHAP integration. Fernandes et al. (2024) reported MAPE of 5.85% in multi-environment maize trials.

Deep Learning:

LSTM & GRU capture temporal patterns in trial data. El-Kenawy et al. (2024) found GNN and LSTM outperformed ML baselines for potato yield. Alsaber et al. (2025) confirmed ANN outperformed SVM and XGBoost in time-series Indian potato yield studies.

Explainability Gap:

Ensemble and DL models are black-box by nature. SHAP (SHapley Additive exPlanations) has emerged as the leading XAI framework for revealing feature importance and GEI interactions in crop models (Yu et al., 2025).

METHODOLOGY

1. EDA & Feature Engineering
Outlier detection, missing values, scaling, Weather & Soil data aggregation (per site/year)

2. Statistical Baseline Development
Linear regression varieties, AMMI, GGE biplot

3. Machine Learning Models Development
RF, KNN, SVM, XGBoost, LightGBM — with hyperparameter tuning via cross-validation

4. Deep Learning Models Development
LSTM & GRU for temporal sequence modelling

5. LOEO Validation
Leave-One-Environment-Out — tests generalisation to unseen sites; prevents data leakage across environments

6. Explainability (XAI)
SHAP values — global & local feature importance, GEI driver analysis

7. Evaluation & Comparison
RMSE, MAPE (<10% desired), R^2, stability metrics across sites

TOOLS & TECHNOLOGIES

Statistical Models
AMMI · GGE Biplot · Regression (Ridge / LASSO / ElasticNet)

Machine Learning
Random Forest · XGBoost · LightGBM · KNN · SVM

Deep Learning
LSTM · GRU

Explainability (XAI)
SHAP — global & local feature attribution

Evaluation Metrics
RMSE · MAPE (<10%) · R^2 · LOEO

Deployment Stack
Streamlit dashboard · FastAPI (Depends on Teagasc Infrastructure)

SYSTEM DEPLOYMENT & MLOps

Each annual harvest cycle generates new variety-environment observations that will be ingested, validated, and used to retrain the model — improving prediction accuracy over time and ensuring the system adapts to emerging climate patterns and new varieties entering the Teagasc potato breeding programme.

◀ CONTINUOUS FEEDBACK LOOP & MODEL RETRAINING ▶

New Data Ingestion	MODEL RETRAINING	PERFORMANCE MONITORING
Breeders flag prediction accuracy vs. actual trial	New season data append to Genovix training set	Validation is performed per season & environment
Observed yield results fed back into system	Models are retrained annually post-harvest	SHAP drift detection — flag shifting GEI drivers
Disagreements logged for model review	LOEO re-validation on updated environment set	Model versioned & audit-logged for Teagasc

EARLY INDICATIONS & EXPECTED OUTCOMES

Literature strongly supports ML/DL superiority over linear statistical models for nonlinear GEI modelling — MAPE below 10% is achievable with ensemble methods.

Expected Outcomes:

- XGBoost / LightGBM predicted to outperform linear statistical baselines
- LSTM/GRU expected to capture sequential seasonal & temporal relationships in MET data
- Explainability analysis with SHAP anticipated to reveal dominant environmental drivers
- **Final deliverable**: An integrated prediction system deployable within Teagasc's potato breeding workflow

SIGNIFICANCE

The Problem It Solves

- Statistical models Treat environment as a category — miss nonlinear Genotype x Environmental Interactions
- Force full multi-year, multi-site trials before a variety can be selected or dropped
- Slow, costly process with limited predictive foresight

The Solution

- ML/DL models predicting marketable yield (kg/ha) at unseen sites — before the trial runs
- LOEO cross-validation — a proven strategy for MET data
- SHAP explanations: not just a prediction, but the reason behind it
- Interactive dashboard — will facilitate **Human In The Loop** interaction

The Benefit to Teagasc

- Earlier, cheaper screening — prioritise varieties before committing to full trials
- Transparent decisions — SHAP reveals why a variety underperforms (e.g. poor dry-matter under low rainfall)
- Decades of trial data put to predictive use for the first time
- Direct support for Ireland's food security & sustainable agriculture goals

REFERENCES

1. Alsaber, A., Satpathi, A., Alsabah, M. and Setiya, P. (2025). Optimizing potato yield predictions in Uttar Pradesh, India: a comparative analysis of machine learning models. Scientific Reports, 15(1). doi:https://doi.org/10.1038/s41598-025-12719-8.
2. El-Kenawy, E.-S.M., Amel Ali Alhussan, Nima Khodadadi, Seyedali Mirjalili and Eid, M.M. (2024). Predicting Potato Crop Yield with Machine Learning and Deep Learning for Sustainable Agriculture. Potato Research. doi:https://doi.org/10.1007/s11540-024-09753-w.
3. Fernandes, I. K.; Vieira, C. C.; Kaio; Fernandes, S. B. Using Machine Learning to Combine Genetic and Environmental Data for Maize Grain Yield Predictions across Multi-Environment Trials. Theoretical and Applied Genetics 2024, 137 (8), 189–189. https://doi.org/10.1007/s00122-024-04687-w.
4. Piekutowska, M. and Gniewko Niedbała (2025). Review of Methods and Models for Potato Yield Prediction. Agriculture, [online] 15(4), pp.367–367. doi:https://doi.org/10.3390/agriculture15040367.
5. Yu, T., Zhang, H., Chen, S., Gao, S., Liu, Z., Wang, J., Crossa, J., Montesinos-López, O.A., Hearne, S. and Li, H. (2025). EXGEP: a framework for predicting genotype-by-environment interactions using ensembles of explainable machine-learning models. Briefings in Bioinformatics, 26(4). doi:https://doi.org/10.1093/bib/bbaf414.

Connect With Me

Industry Partner

Agriculture and Food Development Authority