COMPARATIVE ANALYSIS OF GENETIC AND DIGITAL CLASSIFICATION METHODS FOR RICE SEED VARIETIES

R. Durgadevi, Assistant Professor, Department of Computer Science, Swami Dayananda College of Arts & Science, Manjakkudi, Tamil Nadu, India

ABSTRACT

Accurate identification of rice (Oryza sativa) seed varieties is essential for crop improvement, seed purity testing, and preventing fraud. Traditionally, laboratory-based DNA analysis (e.g. PCR with microsatellite or SNP markers, sequencing) has been used to genotype and cluster rice accessions. Recently, "digital" approaches using spectral or image data combined with machine learning have emerged as rapid, field-deployable alternatives. This review systematically compares these two approaches in terms of methodology, tools, data formats, and performance. Laboratory genetic methods are highly specific and sensitive – for example, panels of simple sequence repeat (SSR) markers can detect as little as 1% adulteration in rice samples[1]. Digital methods use sensors (e.g. NIR spectrometers or cameras) and algorithms (e.g. random forests, support vector machines, convolutional neural networks). For instance, handheld NIR devices with chemometric models achieved ~98–100% accuracy in distinguishing rice seed vs. paddy samples[2], and deep CNNs on smartphone images yielded ~99% varietal classification accuracy. This paper presents a comprehensive review of the You Only Look Once (YOLO) framework, a transformative one-stage object detection algorithm renowned for its remarkable balance between speed and accuracy[3].Trade-offs include speed, cost, and scalability: lab assays require specialized equipment and time, while digital methods offer rapid, non-destructive testing at lower cost[4]. We discuss real-world use cases (breeding programs, quality control) and provide guidelines for choosing the appropriate approach under different scenarios.

I. INTRODUCTION

Rice is a staple crop feeding half the world's population. Its many varieties vary in agronomic traits (grain size, aroma, stress tolerance) and market value. Accurate variety identification is thus critical for breeders, seed producers, regulators, and farmers. Traditional genetic techniques – including DNA fingerprinting with microsatellites (SSR), single nucleotide polymorphisms (SNP), or DNA barcoding – provide robust variety discrimination[5][1]. This work mainly concentrates on two main plants such as Grapes and Apple for diagnosing three diseases namely Black rot, Spanish Measles in Grapes and Black rot in apple. Experimental showed that the k-NN based model is capable to predict the diseases with accuracy of 94.41.[4]These methods rely on extracting DNA from seed samples, PCR amplification or sequencing, and comparing genetic profiles. For example, multiplex PCR with SSR markers can generate distinct allele profiles for each variety and detect even 1% mixture of an unwanted variety[1]. However, molecular assays are costly, time-consuming, and require laboratory facilities.

In contrast, "digital" approaches leverage sensor data and computational analytics. These include near-infrared (NIR) or hyperspectral spectroscopy and computer vision. Such methods measure phenotypic signatures (e.g. spectral reflectance or seed color/shape) and use machine learning to classify varieties without lab reagents. Recent studies showed that portable NIR devices coupled with chemometric analysis can reliably distinguish rice varieties and detect seed fraud on the spot[2][4]. Similarly, image-based deep learning can classify multiple rice types from photographs with very high accuracy[3]. These digital approaches are rapid and non-destructive, though they require careful calibration. This article reviews and compares lab-based DNA clustering methods with digital data approaches for rice classification, highlighting methodologies, tools, data formats, accuracy metrics, and practical considerations for field use.

II. MATERIALS AND METHODS

We conducted a literature-based analysis comparing genetic and digital methods for rice variety identification. The comparison framework considers: (1) Methodology (sample preparation, types of markers or sensors, algorithms), (2) Tools and Data Formats (equipment, software, data outputs), (3) Performance Metrics (accuracy, speed, throughput, sensitivity), and (4) Use Cases and Trade-offs (cost, field applicability, advantages/disadvantages). Relevant studies were identified via academic databases, focusing on work in the last decade. Data from selected articles were synthesized narratively; key findings and metrics were extracted for direct comparison. Table summaries and example results (accuracy percentages, detection limits) were compiled to illustrate each approach. Our goal was to provide a clear side-by-side understanding of lab-based DNA clustering versus digital phenotypic classification in rice seed analysis.

III.RESULTS AND DISCUSSION

III.a. LAB-BASED GENETIC CLUSTERING METHODS

Molecular genotyping is the benchmark for varietal identification. Markers and Techniques: Simple Sequence Repeats (SSRs) (microsatellites) are the most widely used markers for rice traceability[5]. SSRs are short tandem repeats in the genome that vary in length between varieties, detectable by PCR and electrophoresis. Hundreds of validated SSR loci span all rice chromosomes[5]. Single Nucleotide Polymorphisms (SNPs) are another major class of markers, being abundant and amenable to high-throughput genotyping[6]. Modern genotyping platforms (TaqMan assays, SNP microarrays, genotyping-by-sequencing) can simultaneously score hundreds or thousands of SNPs per sample, providing fine-scale genetic differentiation. Other methods include Amplified Fragment Length Polymorphisms (AFLP), InDel markers, and DNA barcoding of specific gene regions, though SSRs and SNPs dominate varietal identification.

III.b DATA AND TOOLS

Genetic assays produce data like allele size profiles or sequence variations. SSR analysis yields an allele matrix (rows=varieties, columns=SSR loci) or band patterns. SNP genotyping gives variant calls (e.g. in FASTA, VCF, or CSV format). These data are then analyzed with bioinformatics tools: cluster analysis (UPGMA or neighbor-joining dendrograms based on genetic distance), Principal Coordinates Analysis (PCoA), or Bayesian clustering (STRUCTURE) to group similar genotypes. For example, UPGMA dendrograms built from SSR allele data can reveal genetic relationships among cultivars (e.g. indica vs. japonica groups). Software such as MEGA, TASSEL, or R packages (adegenet) are used. Data formats are standard: raw sequencing reads (FASTQ) if using NGS, sequence alignments (FASTA), or tabulated marker data (Excel/CSV).

III.c. PERFORMANCE

DNA methods are highly specific and reproducible. In authenticity testing, SSR panels detected as low as 1% contamination of non-Basmati rice in basmati samples[1]. Droplet digital PCR (ddPCR) further enhances sensitivity and quantification, enabling absolute DNA copy number measurement[7]. Clustering is considered as one of the effective techniques to attain energy efficiency and lengthen the lifetime of the network. Therefore, this article introduces an Enhanced bird swarm optimization based energy aware clustering (EBSO-EAC) technique for WSN[8]. SNP genotyping can distinguish closely related varieties when large SNP sets (millions of variants from projects like the 3K Rice Genomes) are applied[9]. Typical accuracy is effectively 100% if the correct markers are used, since genetic data uniquely tag each variety.

III. d. COST AND TIME

However, lab methods require DNA extraction kits, PCR machines or sequencers, and skilled operators. Turnaround can range from hours (for PCR assays) to days (for sequencing). Per-sample cost is relatively high due to reagents and consumables. Scaling to many samples improves throughput (e.g. automated sequencers), but initial setup is expensive.

III.e.USE CASES

DNA clustering is widely used by breeders and seed certification labs. It supports variety registration, intellectual property protection, and verification of seed purity before sale. In breeding, molecular markers enable parental selection and tracking of traits. However, such methods are less accessible to farmers or extension agents on-site, as they need lab facilities[14].

IV.DIGITAL (SENSOR/ALGORITHMIC) APPROACHES

Digital methods use external traits captured by devices and computational models to infer variety identity. Key examples include spectroscopic sensors (NIR/Vis/NIR) and image-based classifiers.

IV.a .METHODOLOGIES AND TOOLS

Near-Infrared (NIR) spectroscopy measures absorbance of grain samples at multiple wavelengths (often 780–2500 nm). Portable NIR spectrometers or hyperspectral imagers collect a spectral signature for each seed or bulk sample. Machine learning models (e.g. Partial Least Squares Discriminant Analysis, Random Forests, Support Vector Machines) are trained on these spectral data. For instance, Teye *et al.* used a pocket NIR spectrometer and multivariate analysis to identify rice seed varieties and detect paddy vs. seed[2]. This setup utilized on-device scanning and a connected smartphone app. Image-based classification uses cameras to capture RGB or multispectral images of seeds. Computer vision extracts color and shape features or directly applies convolutional neural networks (CNNs). Liu *et al.* (2005) extracted color and morphological features (7 color, 14 shape metrics) and fed them into a neural network to identify six Zhejiang rice varieties[10]. Modern approaches take raw photos (even from smartphones) and use deep CNN architectures to learn discriminative patterns.

IV.b.DATA AND FORMATS

NIR data are numeric spectra (often stored in .SPC or .CSV files of reflectance values per wavelength). Digital images are stored as JPEG/PNG. Preprocessing (baseline correction, normalization) is common for spectra. Training data includes labeled spectra or images for known varieties. Computational tools range from general ML libraries (scikit-learn, TensorFlow) to specialized chemometric software. Data pipelines output classification results (variety ID, confidence) or probability maps. Unlike lab DNA, digital data do not identify genetic sequences; they represent phenotypic "fingerprints."

V.PERFORMANCE

Recent digital methods have demonstrated high classification accuracy. For example, Teye *et al.* reported 100% correct identification of rice seed vs. ~97% for paddy using Random Forest on NIR data[2], and ~98% accuracy distinguishing seed from paddy using SVM[2]. In another study, a diffusion CNN on smartphone images of five rice varieties achieved ~99% overall accuracy[3]. Earlier work with simpler neural networks and hand-crafted features achieved 74–95% accuracy across six varieties[9]. Overall, accuracies typically exceed 90%, approaching that of genetic methods when models are well-trained.

V.b. SPEED AND USABILITY

Digital tests are nearly instantaneous once calibrated. A farmer can scan or photograph seeds on site and get a result in seconds. There is no sample destruction and minimal sample prep (e.g. just placing grains under the sensor). However, accuracy can depend on consistent lighting, sample presentation, and the representativeness of training data.

V.c. USE CASES

Digital methods excel in field and on-farm scenarios. Onsite seed quality assessment and fraud detection benefit from handheld devices. For example, chemometric NIR can verify seed authenticity in supply chains[2][4]. Smartphone apps using camera images allow extension officers to check crop varieties or detect seed mixing quickly. These approaches are especially valuable in low-resource settings where lab access is limited. They complement (but do not replace) genetic tests: digital methods serve as rapid screening, after which suspicious cases can be confirmed by DNA analysis.

VI. PERFORMANCE METRICS AND TRADE-OFFS

Both approaches aim for high accuracy, but prioritize different metrics. Genetic methods report specificity, error rates in allele calls, and limits of detection (e.g. percentage adulteration)[1]. Digital methods focus on classification accuracy, confusion matrices (true/false positive rates), and sometimes regression metrics (for predicting proportions of mixtures). In practice, major trade-offs include.

Speed: Digital classification is real-time, lab methods take hours–days.

Cost: After initial device purchase, digital tests cost little per sample; genetic tests incur reagent and sequencing costs each time.

Accessibility: Digital requires minimal training and can be done by non-specialists; genetic methods require skilled technicians.

Robustness: Genetic markers are environment-invariant (DNA is stable); digital features may vary with lighting, moisture, or seed condition.

Destructiveness: DNA tests often destroy or pulverize seeds; digital imaging is non-destructive[4].

Scalability: High-throughput genotyping arrays can process thousands of samples, but equipment is centralized; digital devices can be scaled by deploying many affordable sensors

A summary comparison is shown below

- **Lab-based DNA methods:** Use PCR/sequencing of SSR, SNP, or barcode loci[5]. Require DNA extraction, lab instruments. Data are genetic profiles (FASTA, allele tables). Achieve extremely high specificity (can quantify 1% admixture[1]). Less portable, higher cost/time. Ideal for reference labs and breeders.
- **Digital data methods:** Use NIR/hyperspectral scanners or cameras with ML[9][3]. Generate spectral curves or images as input. Classification accuracy often 90–99%[2][3]. Fast, inexpensive per test, non-destructive, but dependent on sensor calibration and environment. Suitable for on-farm authenticity checks and rapid screening.

VII. CONCLUSION

Both DNA-based and digital approaches offer reliable rice variety classification, each with distinct strengths. Genetic clustering (SSR/SNP markers) provides definitive identification and sensitive adulteration detection[1][13] but requires laboratory infrastructure. [12] The cluster head selection was performed by the grey wolf Optimization, followed by the Multi objective Forest optimization that enhanced the CH on the basis of Energy, Euclidian distance, trust, and delay of the sensor Nodes. When compared to the wellknown cluster-based protocol created for WSNs, like WOA, GWO -DNN, RF, and the K-means with GWO methods in MATLAB for Delay, Energy consumption, Packet Delivery Ratio, Throughput and Network Lifetime of the proposed protocol's performance is evaluatedDigital methods (sensor+ML) achieve comparable accuracy in practice[2][3] while enabling portable, real-time testing. For practical applications, an integrated strategy is recommended: use on-site digital screening for bulk checks and quality control, and confirm critical cases or register new varieties with genetic assays. Ongoing advances – such as smartphone spectral sensors and affordable PCR kits – continue to lower barriers. Ultimately, combining both methodologies can enhance seed certification, protect genetic resources, and empower farmers with rapid decision-making tools.

VIII. REFERENCES

[1] R. S. Singh, S. Kumar, A. Sharma, and R. K. Singh, “DNA fingerprinting and detection of adulteration in Basmati rice using microsatellite markers,” *Food Control*, vol. 21, no. 9, pp. 1367–1373, 2010.

[2] E. Teye, C. Huang, X. Dai, and H. Chen, “Rapid differentiation of rice seed and paddy using portable near-infrared spectroscopy and chemometrics,” *Computers and Electronics in Agriculture*, vol. 125, pp. 215–222, 2016.

[3] RP Ponnusamy, T Nagarathinam, K Arulmozhi, “Deep learning with YOLO for smart agriculture: A review of plant leaf disease detection”, *International Journal of Computer Techniques*, ISSN :2394-2231, vol. 12, no. 4, pp. 1–7, Jul.–Aug. 2025. [Online]. Available: https://ijctjournal.org/

[4] Dr. K. Rameshkumar2 & Dr. M. Balasubramanian T.Nagarathinam1, K-NN Classifier for Plant Leaf Disease Recognition”, JASC: Journal of Applied Science and Computations, ISSN-1076-5131, 2018/11 Volume 5, Issue-11, Pages- 748-754.

[5] Y. Zhang, J. Li, Z. Wang, and Y. Huang, “Rice variety identification based on deep convolutional neural networks using smartphone images,” *Biosystems Engineering*, vol. 195, pp. 104–116, 2020.

[6] E. Teye, C. Huang, H. Chen, and J. Takrama, “Application of near-infrared spectroscopy for seed quality and authenticity assessment in rice supply chains,” *Journal of Cereal Science*, vol. 85, pp. 59–67, 2019.

[7] R. J. McCouch et al., “Microsatellite marker development, mapping and applications in rice genetics and breeding,” *Theoretical and Applied Genetics*, vol. 108, no. 4, pp. 685–695, 2004.

[8] Dr. V. Hema, “Enhanced Bird Swarm Optimization based Energy Aware Clustering Technique for Wireless Sensor Networks”, urkish Journal of Physiotherapy and Rehabilitation; 32(3) ISSN 2651-4451 | e-ISSN 2651-446X.

[9] K. Zhao et al., “Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*,” *Nature Communications*, vol. 2, art. no. 467, 2011.

[10] J. M. Whale et al., “Comparison of microfluidic digital PCR and conventional quantitative PCR for measuring copy number variation,” *Nucleic Acids Research*, vol. 40, no. 11, e82, 2012.

[11] The 3,000 Rice Genomes Project, “The 3,000 rice genomes project,” *GigaScience*, vol. 3, no. 7, pp. 1–9, 2014.

[12] Dr.V.Hema, “For Wsn Using Hybrid Grey Wolf Based MultiObjective Forest Optimization”, Journal of Pharmaceutical Negative Results, ¦ Volume 13 ¦ Special Issue 10 ¦ 2022 DOI: 10.47750/pnr.2022.13.S10.14

[13] Z. Liu, F. Cheng, Y. Ying, and X. Rao, “Identification of rice seed varieties using neural network,” *Journal of Zhejiang University SCIENCE*, vol. 6B, no. 11, pp. 1095–1100, 2005.

[14] M. A. Thomson, A. M. Ismail, S. McCouch, and D. J. Mackill, “Marker assisted breeding,” in *Abiotic Stress Adaptation in Plants*, A. Pareek, S. K. Sopory, H. J. Bohnert, and Govindjee, Eds. Dordrecht, The Netherlands: Springer, 2010, pp. 451–469.