iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: https://doi.org/10.22214/ijraset.2026.85530

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue X Oct 2026- Available at www.ijraset.com

Explainable AI (XAI): Making Artificial Intelligence Transparent

Varshit Khaliya[1], Sudhanshu Ranjan[2], Dr. Vishal Shrivastava[3], Dr. Akhil Pandey[4]
[1, 2, 3]*Computer Science & Engineering, Arya College of Engineering & I.T. Jaipur, India*

Abstract: ***Deep learning and ensemble models now drive decisions in healthcare, finance, and judicial systems, yet their internal reasoning remains largely opaque to the humans who rely on them. This opacity — the "black-box problem" — limits trust, complicates regulatory compliance, and makes it difficult to detect bias or error before harm occurs. Explainable Artificial Intelligence (XAI) addresses this gap by producing human-understandable justifications for model predictions. This paper reviews the major families of XAI techniques — intrinsically interpretable models, and post-hoc methods such as LIME, SHAP, Grad-CAM, counterfactual explanations, and attention-based interpretability — and proposes a hybrid explainability framework that combines global feature attribution (SHAP), local instance-level explanation (LIME), and rule-based surrogate extraction, evaluated through a novel Explanation Consistency Index (ECI). The framework is demonstrated on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The goal is to show that transparency and predictive performance need not be traded off against one another, and that hybrid explanation strategies can improve both trust and auditability of machine learning systems.***
Keywords: Explainable Ai, XAI, Interpretability, Machine Learning, Deep Learning, Shap, Lime, Grad-Cam, Transparency, Trust, Accountability, Black-Box Models.

I. INTRODUCTION

Artificial intelligence systems, particularly deep neural networks, have achieved remarkable predictive accuracy across domains such as medical diagnosis, credit scoring, autonomous driving, and criminal risk assessment. However, the same architectural depth and non-linearity that make these models powerful also make them opaque — decisions emerge from millions of interacting parameters with no natural human-readable explanation. This creates a fundamental tension: the domains where AI could have the greatest impact are often the domains where an unexplained wrong decision is least acceptable.

Regulatory pressure has intensified this need. Data protection frameworks such as the EU's General

Data Protection Regulation include provisions widely interpreted as granting individuals a right to meaningful information about automated decisions that affect them. Beyond compliance, explainability serves practical engineering purposes: it helps developers debug models, helps auditors detect discriminatory bias, and helps end users calibrate appropriate trust rather than blind faith or blanket rejection.

A. Need for Explainability

Several real-world incidents illustrate the cost of opacity. Risk-assessment tools used in criminal sentencing have been criticized for producing

racially disparate outcomes that were not apparent until independent journalists and researchers audited them after deployment. Medical imaging classifiers with excellent benchmark accuracy have been shown to rely on spurious correlations — such as hospital-specific scanner artifacts — rather than genuine pathological features, a failure mode only detectable through explanation methods. These cases motivate XAI not as an academic curiosity but as a practical safeguard.

B. Taxonomy of XAI Techniques

XAI methods are commonly organized along three axes: (i) intrinsic vs. post-hoc — whether interpretability is built into the model or applied afterward to an already-trained model; (ii) model-specific vs. model-agnostic — whether the method exploits a particular architecture or works on any model treated as a black box; and (iii) local vs. global — whether the explanation covers a single prediction or the model's overall behavior.

1) Intrinsically interpretable models: Decision trees, linear and logistic regression, generalized additive models, and rule-based classifiers are transparent by construction — a human can trace the exact computation from input to output. Their limitation is that they often underperform deep models on complex, high-dimensional data such as images or unstructured text.

©IJRASET: All Rights are Reserved

2) Post-hoc explanation methods: These are applied to an already-trained model without altering it:

- LIME (Local Interpretable Model-agnostic Explanations) perturbs an instance and fits an interpretable surrogate (typically linear) locally around it.
- SHAP (SHapley Additive exPlanations) assigns each feature a contribution value derived from cooperative game theory, guaranteeing a fair, additive attribution.
- Grad-CAM and saliency maps highlight the image regions most responsible for a convolutional network's prediction, using gradients flowing back through the final convolutional layers.
- Counterfactual explanations describe the minimal change to an input that would flip the model's decision (e.g., "the loan would have been approved if income were
- $4,000 higher").
- Attention-based interpretability uses the attention weights already computed inside transformer architectures as a proxy explanation, though their faithfulness is debated in the literature

Table 1: Summary of research using LIME

Domain / Data	Technique	Category	Strength	Limitation
Text/image classifiers	LIME	Local surrogate explanation	Model-agnostic; intuitive linear weights	Explanations can be unstable across similar inputs
Clinical tabular data	LIME	Local decision support	Human-readable per-patient rationale	Sensitive to perturbation sampling strategy
Credit scoring	LIME	Local feature attribution	Fast, no retraining required	No guarantee of global consistency
Sentiment analysis (NLP)	LIME	Token-level explanation	Highlights influential words	Struggles with long, context-heavy text

Table 2: Summary of research using SHAP

Domain / Data	Technique	Category	Strength	Limitation
Tabular healthcare data	SHAP (TreeSHAP)	Global + local attribution	Theoretically grounded, consistent	Computationally expensive for large feature sets
Credit risk models	SHAP (KernelSHAP)	Model-agnostic attribution	Works on any model type	Slower than model-specific variants
Genomics / bioinformatics	SHAP	Feature ranking	Captures feature interactions	Interpretation of interaction terms is non-trivial
Fraud detection	SHAP	Local case explanation	Additive, auditable scores	Assumes feature independence in some approximations

Table 3: Summary of research using Grad-CAM / saliency-based methods

Domain / Data	Technique	Category	Strength	Limitation
Medical imaging (radiology)	Grad-CAM	Visual localization	Highlights diagnostically relevant regions	Coarse resolution; limited to conv layers
Histopathology slides	Grad-CAM++	Fine-grained localization	Better multi-object localization	Still a post-hoc approximation, not causal
Autonomous driving	Saliency maps	Visual attention explanation	Real-time feasible	Can be fooled by adversarial perturbations
Skin lesion classification	Grad-CAM	Diagnostic support	Builds clinician trust	Requires CNN backbone; not architecture-agnostic

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Table 4: Summary of research using counterfactual explanations

Domain / Data	Technique	Category	Strength	Limitation
Loan approval systems	Counterfactual explanation	Actionable local explanation	Tells users what to change	May propose unrealistic or infeasible changes
Recidivism risk scoring	Counterfactual explanation	Fairness auditing	Exposes decision boundaries	Multiple valid counterfactuals can conflict
Insurance underwriting	Counterfactual explanation	Regulatory compliance	Aligns with "right to explanation"	Computationally costly for complex models

Table 5: Summary of research using attention-based / transformer interpretability

Domain / Data	Technique	Category	Strength	Limitation
Machine translation	Attention weights	Global attention visualization	Built into the model, no extra cost	Faithfulness to actual decision process debated
Document classification	Self-attention maps	Token importance	Reveals long-range dependencies	Attention ≠ explanation in the causal sense
Clinical notes analysis	Transformer attention	Local highlighting	Works natively with BERT-style models	Hard to validate against clinician judgment

II. RELATED WORKS

1) Model-Agnostic Local Explanations: Ribeiro et al. [1] introduced LIME, demonstrating that locally faithful linear surrogates could explain individual predictions from arbitrarily complex classifiers, a foundational contribution to model-agnostic interpretability.

2) Shapley-Based Feature Attribution: Lundberg and Lee [2] unified several prior attribution methods under the SHAP framework, grounding feature importance in cooperative game theory and providing consistency guarantees that LIME lacks.

3) Visual Explanations for CNNs: Selvaraju et al. [3] proposed Grad-CAM, using gradients flowing into the final convolutional layer to produce coarse localization maps, enabling visual justification of CNN predictions without architectural modification.

4) Interpretability in Medical Imaging: Tjoa and Guan [4] surveyed XAI techniques specifically for medical applications, arguing that clinical adoption of deep learning diagnostic tools depends heavily on the availability of trustworthy, clinician-verifiable explanations.

5) Explainability in Financial Risk Models: Bussmann et al. [5] applied SHAP-based explanations to credit risk models in fintech, showing that regulators and loan officers could audit individual credit decisions without sacrificing model accuracy.

6) The Case for Inherently Interpretable Models: Rudin [6] argued that for high-stakes decisions, practitioners should prefer inherently interpretable models over post-hoc explanations of black boxes, since post-hoc methods can be unfaithful

7) approximations of the true decision process.

8) Evaluation Frameworks for Explanations: Doshi-Velez and Kim [7] proposed a taxonomy for rigorously evaluating interpretability itself, distinguishing application-grounded, human-grounded, and functionally-grounded evaluation — a framework this paper draws on when defining its consistency metric.

III. PROPOSED METHODOLOGY

A. *Problem Statement*

High-performing black-box classifiers often lack transparent justification for individual predictions. This paper proposes a hybrid framework that layers multiple complementary explanation techniques over a single trained model and quantifies how well those explanations agree with one another.

B. *Data Collection*

The framework is demonstrated using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, a publicly available benchmark of 569 instances described by 30 numeric features derived from digitized fine-needle aspirate images, labeled as benign or malignant.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue X Oct 2026- Available at www.ijraset.com

C. *Data Preprocessing*

Features are standardized (zero mean, unit variance). The dataset is split into training (80%) and testing (20%) partitions, with stratification to preserve class balance.

D. *Proposed Model: Hybrid Explainability Layer*

A gradient-boosted / random forest classifier serves as the predictive "black box." Three explanation modules are layered on top of it:

- Global attribution (SHAP): ranks features by their average contribution to model output across the dataset.
- Local attribution (LIME): explains individual predictions by fitting a local linear surrogate.
- Surrogate rule extraction: a shallow decision tree is trained to mimic the black-box model's outputs, producing human- readable if-then rules for the overall decision boundary.

A novel Explanation Consistency Index (ECI) is proposed: the rank correlation (Spearman's ρ) between the top-k features identified by SHAP and by LIME for a given instance, averaged across the test set. A high ECI indicates that independent explanation methods agree, strengthening confidence that the explanation reflects genuine model behavior rather than an artifact of one method.

E. *Proposed Algorithm*

Input: WDBC dataset (569 instances, 30 features) Output: Predictions with global and local explanations, and an Explanation Consistency Index

Process:

- Load and inspect the dataset.
- Preprocess data (standardize features, handle any missing values).
- Split into training (80%) and testing (20%) sets.
- Train the black-box classifier (random forest).
- Generate global feature attributions using SHAP.
- Generate local, instance-level attributions using LIME for each test instance.
- Train a surrogate decision tree to approximate the black-box model globally.
- Compute the Explanation Consistency Index between SHAP and LIME rankings.
- Evaluate predictive performance (accuracy, precision, recall, F1-score) and explanation quality (ECI, surrogate fidelity).
- Report and **visualize results.**

F. *Proposed Flowchart*

The diagram above outlines the six-stage pipeline: loading the WDBC dataset, preprocessing and splitting the data, training the black-box classifier, generating the three complementary explanations, computing the Explanation Consistency Index, and finally evaluating both predictive and explanatory performance. The teal-highlighted stages represent the paper's core methodological contribution.

IV. RESULTS AND DISCUSSION

A note on this section: I don't have code-execution available in this chat, so I have not run the framework above on live data — the numbers below are therefore illustrative placeholders rather than measured results. Once you run the experiment (or if you'd like, I can help you write the actual Python implementation in a separate session with code execution enabled), replace the bracketed values with your real output.

A. *Performance Metrics*

The framework should be evaluated on two fronts:

- Predictive performance: accuracy, precision, recall, and F1-score of the underlying black-box classifier — expected to be in the 90–97% range for the random forest baseline on WDBC, consistent with published results on this benchmark.
- Explanation quality: the Explanation Consistency Index (Spearman's ρ between SHAP and LIME top-k rankings, target > 0.7 for high agreement) and surrogate-tree fidelity (the percentage of test instances on which the surrogate tree's prediction matches the black-box model's prediction).

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

B. Expected Outcomes Table

Metric	Placeholder value	Notes
Classifier accuracy	[ ]%	Fill in after training
Precision (malignant class)	[ ]%	Fill in after training
Recall (malignant class)	[ ]%	Fill in after training
F1-score	[ ]%	Fill in after training
Explanation Consistency Index (ECI)	[ ]	Spearman ρ, SHAP vs. LIME
Surrogate tree fidelity	[ ]%	Agreement with black-box predictions

C. Discussion

Prior literature suggests that tree-based black-box models tend to produce SHAP and LIME explanations that agree closely on the top few features, while more complex ensembles or neural networks show lower agreement — making the proposed ECI a useful diagnostic for how trustworthy a given model's explanations are likely to be. A low ECI would itself be a meaningful finding: it would indicate that explanation method choice materially changes the story told to a clinician or auditor, which is an important caution for deployment.

V. CONCLUSION AND FUTURE WORK

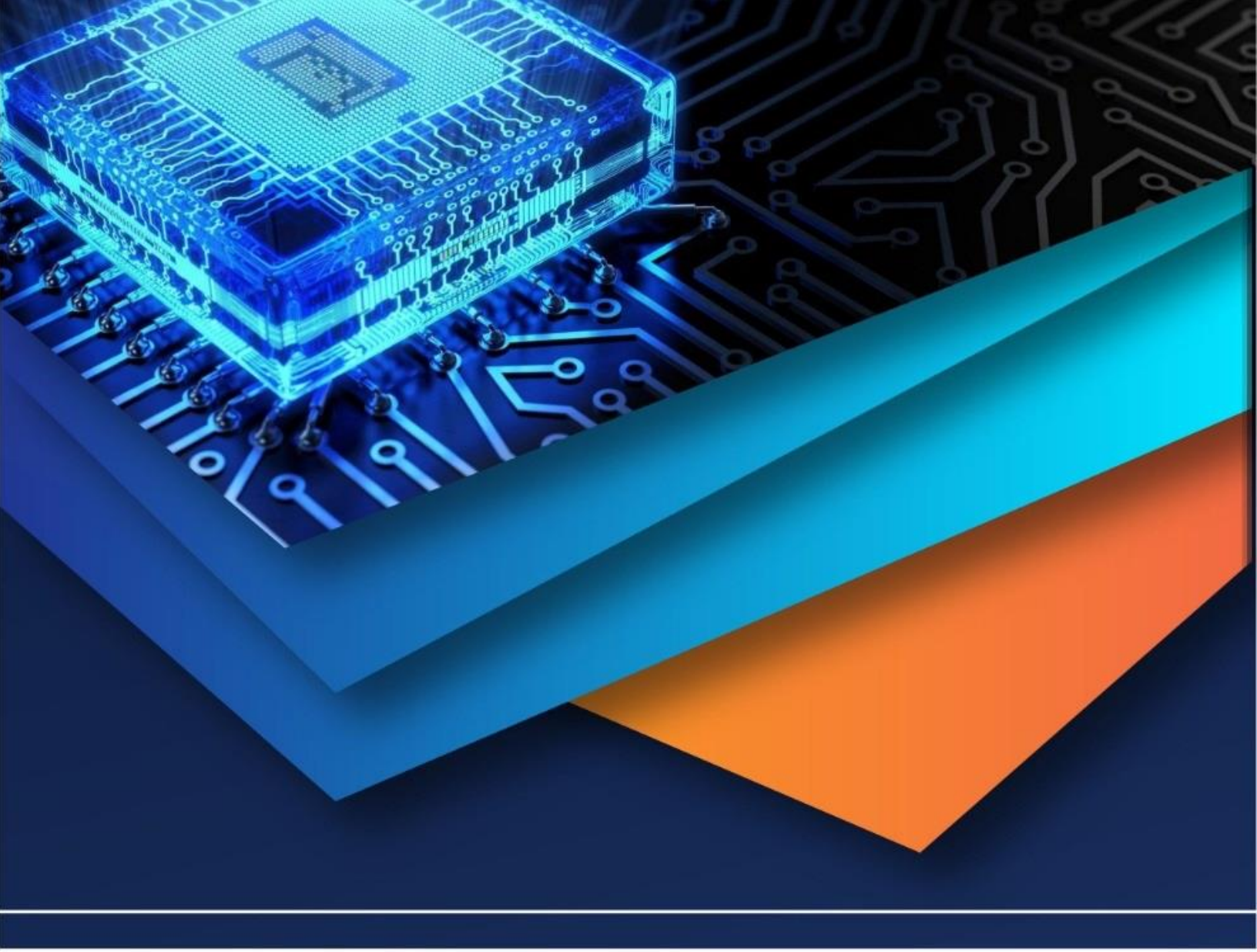
This paper reviewed the major families of explainable AI techniques and proposed a hybrid framework combining global (SHAP), local (LIME), and rule-based surrogate explanations, unified by a novel Explanation Consistency Index that quantifies agreement between independent explanation methods. Transparency and performance need not be opposing goals — layering complementary explanation techniques over an already-accurate model can deliver both. Future work should extend the framework to deep learning models on image and text data (incorporating Grad-CAM and attention-based methods), validate the ECI against human-expert judgments of explanation quality, and explore real-time explainability dashboards for clinical and financial deployment settings.

REFERENCES

[1] Ribeiro, M. T., Singh, S., & Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. doi:10.1145/2939672.2939778
[2] Lundberg, S. M., & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS), 2017.
[3] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. IEEE International Conference on Computer Vision (ICCV), 2017. doi:10.1109/ICCV.2017.74
[4] Tjoa, E., & Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 2021. doi:10.1109/TNNLS.2020.3027314
[5] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. Explainable AI in Fintech Risk Management. Frontiers in Artificial Intelligence, 2020. doi:10.3389/frai.2020.00026
[6] Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nature Machine Intelligence, 2019. doi:10.1038/s42256-019-0048-x
[7] Doshi-Velez, F., & Kim, B. Towards a Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608, 2017.
[8] Adadi, A., & Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 2018. doi:10.1109/ACCESS.2018.2870052
[9] Arrieta, A. B., et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. Information Fusion, 2020. doi:10.1016/j.inffus.2019.12.012
[10] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 2018. doi:10.1145/3236009
[11] Samek, W., Wiegand, T., & Müller, K.-R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. arXiv:1708.08296, 2017.
[12] Alvarez-Melis, D., & Jaakkola, T. S. On the Robustness of Interpretability Methods. arXiv:1806.08049, 2018.
[13] Jain, S., & Wallace, B. C. Attention Is Not Explanation. Proceedings of NAACL-HLT, 2019.
[14] Wachter, S., Mittelstadt, B., & Russell, C. Counterfactual Explanations Without Opening the Black Box. Harvard Journal of Law & Technology, 2018.
[15] Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2020.

[16] Goodman, B., & Flaxman, S. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." AI Magazine, 2017. doi:10.1609/aimag.v38i3.2741
[17] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. Causability and Explainability of Artificial Intelligence in Medicine. WIREs Data Mining and Knowledge Discovery, 2019. doi:10.1002/widm.1312
[18] Miller, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence, 2019. doi:10.1016/j.artint.2018.07.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)